ログのカテゴリー変数に対する ダミー変数と項目マッピングを用いた行列変換処理手法

Matrix Transformation Processing Method Using Dummy Variables and Item Mapping for Log Categorical Variables

輪島　幸治† Aminanto Muhamad Erza† 班　涛† 伊沢　亮一† 高橋　健志† 井上　大介†

† 国立研究開発法人 情報通信研究機構
〒 184–8795 東京都小金井市貫井北町 4-2-1
E-mail: †{wajimak,aminanto,bantao,isawa,takeshi_takahashi,dai}@nict.go.jp

あらまし　近年，情報通信技術の発展により，多くの情報アクセス機器が登場した．情報アクセス機器には，利用者に対するセキュリティ対策で，不正通信通知機能がある．その一方で，情報アクセス機器が増加することは情報通信機器の数だけ，出力ログ数が増加することに相当する．出力ログ数の増加は，分析作業における作業が増加することから，分析者のデータクレンジングのコスト増加となる．本研究では，IDS のダンプデータを用いて, 大量のアラートを効率的に行列データに変換処理を行う手法を提案する．IDS のダンプデータに対して，提案手法を適用し，既存の機械学習手法で評価した結果，評価用行列として平易に作成できることや，項目評価や特徴量変換に適用できることで有用性が明らかになった．結果を報告する．

キーワード　アラートスクリーニング，データクレンジング，カテゴリー変数，ダミー変数化，Random Forest

1 はじめに

近年，情報通信技術の発展により，多様な IT サービスおよび情報アクセス機器が増加した．電気通信回線を用いた Web サービス，オンプレミスなど多種多様なシステムが提供されており [1]，分析するログは増加してきている．ログを収集・分析するシステムなども増加してきており，収集したデータを，リッチな UI で分析する製品サービスなども登場してきている [2]．

一方で，情報アクセス機器に対する侵入や攻撃，セキュリティ対策に対する新たな回避テクニックなど，不正な通信に起因したインシデントも増加してきている．インシデントとなるセキュリティの脅威は，攻撃の種類も方法も多様化している [1]．セキュリティの脅威に対する対策には，不正通信の早期検知を目的に，アラートログの分析が重要視されている．

ここで，各機器に対するセキュリティ対策が増加した場合，担当者に通知されるアラート数も増加する．システムの動作や設計の問題でアラート数が多いだけのことは，安心感の低下であり，知識がないユーザにとっては，心理的な問題要因にもなる [2]．ゆえに，担当者は多数のアラートから，早期に対応すべきアラートを切り分けなければならない．

重要なアラートの切り分けを目的とする作業の一つにアラートスクリーニングがある．そもそも，スクリーニングとは，膨大な量のインフラを定期的に診断した結果やログから，要因となるべき要素を絞り込み，抽出する作業である [3]．しかし，ログ収集機器で出力ログは異なる [3] [4]．また，適切なスクリーニングは，判別すべき担当者が，有識者である必要があり，非常にコストがかかる．加えて，各特徴を表すベクトルが非常に大きくなった場合，次元の呪い [4] といった異なる問題を引き起こす場合がある．ゆえに，効率的に分析するために最重要となるのは，対象のデータを分析が行えるデータに変換処理することである．現状では，各分析者がデータクレンジング処理を用いて行列データ化していることから，一意的な手法は提案されていない．したがって，データ分析におけるクレンジング処理は，分析者の大きな負担であることが課題となっている．

そこで，本研究では，アラートスクリーニングにおけるデータクレンジング処理に着目して，出力ログを効率的に行列データ化する手法を提案する．提案手法では，ログの抽出レイヤーにかかわらず，ログに含まれるカテゴリー変数から，ダミー変数化を用いて，分析可能な数値データの行列に変換処理する．

1 : Oracle - Documentation : https://docs.oracle.com/en/
2 : SAP - Portfolio Categories : https://www.sap.com/products.html
3 : Cisco - All Products Support : https://www.cisco.com/c/en/us/support/all-products.html
4 : 次元数が大きい場合に，解析結果の精度や安定性に影響や困難が生じること．表現次元は高いが，本質次元が低い場合などもある [5].

また，提案手法は，項目の判別をログのユニーク数に基づいて必要な項目を判別した．ゆえに，必要以上に大きな行列データを処理せずに，変換処理が行える．変換処理が最適化されれば，処理されていない多くのデータを処理でき，分析者が本来行う作業に注力できる．結果，データクレンジング処理における分析コストの低下においても期待できる．本研究では，提案手法における実装をすべて既存のライブラリで実現した．既存のライブラリを用いることで，一意的な結果を得ることが期待でき，分析者の実装コストも低下することが期待できる．提案手法で得られた行列データを特徴選択手法および特徴変換を用いた機械学習手法で評価した．結果を報告する．

2 関 連 研 究

2.1 データ・マイニング

情報通信社会の発展で，ソーシャルメディアを始めとした，多様な情報発信が行われてきており [6]，多くの分析が行われている．データにおける分析では，データベースやログに含まれる有益なパターンやルールを基に，有効性を評価・判別する．情報発信が行われるとデータベースに，情報が蓄積される．既に発信済みの文献公開 URL だけでも多数存在している [5] [6] [7] [8] [9] [10]．ゆえに，データベースには大量のデータが存在している．ここで，コンピューター上のデータベースに蓄積された大量のデータから，鉱脈を探り当てるように繰り返し，相関性を探し出すことをデータ・マイニングと呼ぶ [7]．一方で，有益な知見を発見するためにはデータを処理できるデータに変換する必要がある．データマイニングの目的は，データから，新たな有益なパターンやルールを作成することである [8]．

データマイニングでは，多様な取り組みが行われている [11]．一般的には有益なパターンやルールは，人が解釈・理解しやすい結果とすべきである．それゆえ，データを処理可能なデータに変換することが重要である．処理が行えるデータに変換することができれば，マーケティング担当者や広告担当者などが業務情報に使用するデータウェアハウス [7] などへの応用も期待できる．大量のデータから有益な知見を発見する研究は IoT 分野においても，発展が期待されている [9]．

2.2 セキュリティ分野におけるデータ解析

セキュリティ分野は広範な研究分野である．ペネトレーションテスト [12]などの研究も行われており，製品化 [13]も多くなされている分野である．本研究では，IDS などの通信データを用いた侵入検知や不正通信に関して概説する．

セキュリティ分野における研究では，アラートにおける攻撃シナリオ，攻撃フォーカス認識などを定義して評価する特徴的な評価を行う方法などもある [10]．侵入検知やアラートに関する研究は多くの研究が行われている [11] [12] [13]．不正通信における悪意のある動作の検知や侵入検知では，ミスユース検出と異常検知に大別される．不正通信を識別して，セキュリティ対策を行うという目的は同じだが，方法が異なる．ミスユース検出では，不正アクセスのパターンが登録されているデータベースと照合して，判別が行われている．不正アクセスのパターンが明らかであれば，検知できる．対して，異常検知の目的としては，以前は未知であった攻撃パターンを明らかにする方法である．異常検知における異常とは，正常なパターンに該当しないパターンであると定義している [14]．ミスユース検出のメリットとしては，不正アクセスのパターンが明らかであれば，検知に注力できる．一方で，異常値検出においては，正常な領域境界の近くにある観測があいまいである場合や悪意のあるアクションの結果である場合，異常な観察を正常に見せるように適応する．したがって，課題もあるとされている [14]．盛んな研究が行われているが，膨大な量のログから，要因となるべき要素を絞り込み，分析することが必要である．ゆえに，最重要課題とされるべきは，対象のデータを平易かつ一意的に処理できるデータに変換することである．

2.3 データ・クレンジング

対象のデータを分析可能なデータ形式に変換する作業は，データクレンジング処理と呼ばれる．データクレンジング処理は，分野に応じて異なる．日本語の自然言語処理分野においては，自然言語の文字列を分割する単語分割処理がデータクレンジング処理である [15]．具体的には，文字列を一定の規則にしたがった区切りに分割する形態素解析や分かち書きを行うことである．データクレンジング処理では，一意的な処理を行う必要があるため，形態素解析器と呼ばれるアプリケーションが用いられる [14]．また，音声処理分野においても，自然言語を対象としていることから，類似の手法が用いられているが，音声の書き起こしという追加の処理後に文字列として処理される．音声の書き起こしに関しては，人手による書き起こしだけでなく，一意的な処理を目的に，自動書き起こしのアプリケーション [15]も用いられている．ゆえに，分析においては，適切なデータクレンジングを行い，処理できるデータに変換する必要がある．データクレンジング処理された結果に対して，トピックモデル [16] をはじめとした機械学習や統計解析などのアルゴリズムを適用することで，データの中から有益な知見を得ることが期待できる．また，目的変数があれば，多変量解析などへの応用も期待できる．

3 提 案 手 法

3.1 概　　要

情報アクセス機器に対する不正な通信や攻撃の検知においては，適切な判別することが求められている．しかし現状，人手で大量のアラートを判別することは，困難である．

---

5 : J-STAGE : https://www.jstage.jst.go.jp/browse/-char/ja
6 : 情報学広場 : https://ipsj.ixsq.nii.ac.jp/ej/
7 : IEICE Publications Search : https://www.ieice.org/publications/search/
8 : ACM Digital Library : https://dl.acm.org/dl.cfm
9 : IEEE Xplore Digital Library : https://ieeexplore.ieee.org/Xplore/home.jsp
10 : 特許情報プラットフォーム (J-PlatPat) : https://www.j-platpat.inpit.go.jp/
11 : Kaggle : https://www.kaggle.com/
12 : 脆弱性診断の一つ．事前にヒアリング・準備され，ソーシャルエンジニアリング手法も用いられる場合もある．
13 : FireEye - Penetration Testing : https://www.fireeye.com/services/penetration-testing.html
14 : MeCab : https://taku910.github.io/mecab/
15 : Julius : https://julius.osdn.jp/

大量のアラートを自動分析して，スクリーニングする作業が必要とされているが，分析を行うためのデータクレンジングが，大きな課題となっている．本研究では，IDS におけるアラートデータを対象に，平易かつ一意的に処理可能な形式に変換する処理手法を提案する．平易かつ一意的に処理できることで，対象言語に関わらず，多様なログデータの処理が期待できる．本研究における提案手法のステップを下記に示す．

提案手法のステップ

Step.1 解析処理

解析処理として，対象となるデータを解析処理し，ファイル型の変換を行う．*json* ファイルから *csv* ファイルに変換する処理などに相当する．節 3.2 にて示す．

Step.2 ダミー変数化と項目マッピング

ファイル型が変換された入力に対して，値に基づいて行列化する処理を行うことに相当する．行列化処理では，ダミー変数化を用いて，値を数値化した後，項目マッピング処理で，使用項目を判別する．作成された各列のダミー変数行列を横方向に結合し，評価対象となる節 3.3 にて示す．

Step.3 観測行列の処理方法

Step.2 で得られた観測行列 Y を特徴量変換手法を用いて変換特徴量を算出する．節 3.4 にて示す．

本研究では，ログデータの値に着目して，実装を平易かつ一意的にダミー変数に変換することを目的に，Pandas の *get_dummies* メソッドを用いた [16]．

3.2 解 析 処 理

解析対象だが，実データは企業のコンプライアンスや情報セキュリティの影響を懸念し，一般には公開されないことが多い．ログ出力機器のレイヤ [4] や言語でも異なる．商品工場における生産機器など製品製造機器のログ分析は，企業のサステナビリティー [17] の観点から，盛んに分析が行われている．近年は，持続可能性への配慮が世界的潮流となっていることから [18]，産業の新陳代謝を高めることや [17]，研究開発の活性化にも効果的である [18]．そこで，まず本節では，提案手法を広範なログに適用できるよう，具体的に，公開されているログを示して，説明する．本研究で用いる具体例は，カーネギーメロン大学が提供しているデータセット「Enron Email Dataset」(2019/12/15 : Last confirmation date) [19] [20]を用いる．Enron[21]の Email Dataset における処理は，サクラエディタ [22] の文字列検索メソッドで Message-ID[23]が含まれるファイル一覧を取得して，処理した．データ件数は 521,449 件である．データセットを解析して，カテゴリデータを含むログデータに変換処理を実施した．

変換結果を，表 1 に示す [24]．本研究では，各項目の値抽出は，前方一致検索で抽出して，項目データに解析処理した．

表 1　システムが出力するログデータの例 (Enron Email Dataset の場合)

項番	項目名	データ型	unique	使用想定項目
1	Message-ID	text	515,508	○ (加工)
2	Date	text	224,072	○ (加工)
3	From	text	20,314	○ (加工)
4	To	text	51,033	○ (加工)
5	Subject	text	-	
6	Cc	text	27,820	○ (加工)
7	Mime-Version	**int & float**	2	○
8	Content-Type	text	3	○
9	Content-Transfer-Encoding	text	4	○
10	Bcc	text	27,820	○ (加工)
11	X-From	text	27,969	
12	X-To	text	73,507	
13	X-cc	text	33,686	
14	X-bcc	text	133	○
15	X-Folder	text	430	○
16	X-Origin	text	260	○
17	X-FileName	text	5336	○
18	E-Mail Body	text	-	○

表 1 から明らかなようにログデータの場合は，抽出結果はテキスト項目が多いことがわかる．しかし実際に，ログデータに対して，機械学習を適用して，評価する場合は，数値データ化する必要がある．Message-ID 項目などは直接は使用しづらく，Message-ID におけるドメイン部分を抽出して集計するなど，各項目の加工処理も重要である．また，X-FileName 項目などは text 項目であり，“vkamins.nsf” [25] のようなファイル名が値である．表 1 をグループ化した場合の例を表 2 に示す．

表 2　入力が想定される行列データの例

項番	データ型	実データの例	対応項目
1	int & float	10, 20, 30 (値)	(入力想定)
2	int	1, 2, 3 (頻度)	(入力想定)
3	int	1, 2, 3 (カテゴリ)	7
4	text	’alpha’, ’beta’, ’gamma’	17
5	text (Trim)	’6bit’, ’7bit’, ’8bit’	2,9
6	text (Process)	’critical.notice@enron.com’	1,3,4,5,6,8, 10,11,12,13, 14,15,16,18

具体的に使用する項目だが，表 2 における，項番 1 や項番 2 は，Enron Email Dataset における項目にない項目だが，入力が想定される項目であり，値に応じて直接使用すれば良い．項番 5 は，必要箇所以外を除去 (bit を削除) して，値を数値あるいは数値のカテゴリー変数に変換してから使用する方法，項番 6 は，必要箇所のみを処理する方法 (@を集計して数値データ化) などが想定される．本研究では，項目 3 および項目 4 を用いて，提案手法における項目値の変換方法を示す．

16 : pandas-dev - core: https://github.com/pandas-dev/pandas

17 : サステナビリティーは，持続可能性のことである．ESG の課題という広範な社会の目的を達成するために国連責任投資原則（PRI）なども定められている．

18 : 日経 ESG : https://project.nikkeibp.co.jp/ESG/

19 : Enron Email Dataset : http://www.cs.cmu.edu/~enron/

20 : Enron Email Dataset を用いた先行研究もあり，他の URL においてもデータセットが提供されている．必要に応じて，参照頂きたい．

21 : かつてアメリカに存在した企業である．2000 年度年間売上高 1,110 億ドル（全米第 7 位），2001 年度社員数 21,000 名の全米で有数の大企業であった．粉飾決算が影響で倒産した．粉飾決算などはストックオプション制度などと関連が深いが，アメリカ社会における文化的な内容は文献 [19] などを参照頂きたい．

22 : サクラエディタ : https://sakura-editor.github.io/

23 : E-Mail で用いられる識別子の一つ．

24 : 本研究では，*json* ファイルから *csv* ファイルに変換する箇所に相当する．

25 : (参考).nsf は，“Notes(Lotus Notes)” と言うアプリケーションのファイルである．Lotus Notes/Domino の歴史 :
https://www.ibm.com/developerworks/jp/lotus/library/ls-NDHistory/

3.3 ダミー変数化と項目マッピング

ダミー変数化処理は入力行列を対象にダミー変数を用いた数値行列に変換する処理である．表 2 における項目 3 および項目 4 を用いて，説明する．ある 2 種類の項目の文字列リスト「“alpha,1”,“beta,2”,“gamma,3”,“alpha,1”,“gamma,3”,“beta,2”」が与えられた場合に，6 × 6 の行列に変換した例を式 (1) に示す．入力文字から，文字の種別は 2 種類であり，各種別におけるアイテムには，3 種類の文字列があることがわかる．

$$\begin{bmatrix} alpha & beta & gamma & “1” & “2” & “3” \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{1}$$

結果から，表 1 で示した通り，ダミー変数化した場合，種別が 2 列，各種別における値の列が各 3 列に変換されていることがわかる．ゆえに，6 × 6 の行列に変換される．全てのデータを行列変換することは，理論上は容易である．しかし，一意的な値が多い項目をダミー変数化すると，値の数だけ列が作成されることとなる．ゆえに，時間が経過するにしたがって，非常に大きなベクトルが作成されることとなる．これは与えられる項目値の種類が多ければ多いほど作成されるダミー変数の列数も増加することに相当する．具体的には，IP アドレス項目の場合，列数は，IP アドレスの数だけ増加する．したがって，計算量が多く，処理が十分に行えない場合もあり，適切な処理が必要とされている [26]．[27] そこで，本研究では値のユニーク値 (列数となるべき値の種類数) にスレッショルドを設定した．本研究では，スレッショルドを超過した場合に，判定ラベルを項目に設定した．機械学習の入力に変換行列の結果をマッピングすることから，この処理を本研究では，項目マッピング処理と呼ぶ．判定ラベル (L0 もしくは L1)，入力データにおける入力項目数を $(i = 1, ..., N)$，入力データを $(x_1, x_2, ...x_N)$，項目における値のユニーク値を算出する関数を $unique(x_N)$ とした場合を定式化した項目マッピング処理を式 (2) および式 (3) に示す．

$$X = \left[x_1, x_2 \cdots x_N\right] \tag{2}$$

$$x_N = \begin{cases} unique(x_i) < \text{Threshold}, & x_i \in Set \quad L0 \\ unique(x_i) > \text{Threshold}, & x_i \in Set \quad L1 \end{cases} \tag{3}$$

式 (3) から，$Threshold$ 未満の場合 $L0$ が設定される．超過した場合 $L1$ が設定される．本研究では，判定ラベルが L0 にラベル付けされたダミー変数の各処理対象項目と直接使用する項目を横方向に結合させた行列を分析を行う観測行列 Y とした．また，式 (3) のスレッショルドには $Threshold$ を設定した．

26：(参考) 本研究の範囲対象外だが，NAT や IP マスカレードなどを除き，IP アドレスはユニークな場合が多い．ゆえに，IP アドレスからレピュテーションスコアを算出して，スコア化して評価するなど，変換処理には用いずに，異なる値変換する場合も想定される．

27：(参考)IP アドレスの処理方法には，IP アドレスやドメインが，ブラックリスト (DBL) に登録されていないかなどをスキャンして，識別やフィルタリングする処理などもある．
The Spamhaus Project：https://www.spamhaus.org/

3.4 観測行列の処理方法

本節では，ダミー変数に変換された行列の使用方法に関して述べる．本研究では，特徴量変換を用いた変換特徴量の分布で，ダミー変数化の妥当性や，アラート評価への効果を明らかにする．特徴量変換におけるアルゴリズムは，多数のアルゴリズムが提案されている [20] [21] [22] [23] [24] [25] また，各手法の拡張や応用も行われている [26]．本節では，説明のため乖離度規準や損失関数など多様な応用が行える非負値行列因子分解 (NMF：Non-Negative Matrix Factorization) に基づいて説明を行う [20] [28]．NMF は，与えられた観測行列 Y を変換特徴量と変換特徴量に対する寄与率の行列に分解するアルゴリズムである．以後，ログの特徴量である観測ベクトルを並べた行列を観測行列をみなして表記する．まず，観測行列を Y とおき，出力ログ数 $(i = 1, ..., N)$，特徴量の次元数に相当するダミー変数化した出力ログの項目数の次元数 $(j = 1, ..., K)$ とする．この場合，構成される観測行列は，N 行 K 列の長方行列である [29]．NMF を適用する場合，観測行列 Y の次元数 K よりも，変換特徴量となる基底数 $M(m = 1, ..., M)$ を小さく設定して，観測行列 Y を，低ランクの行列の積で近似することに相当する．

2 つの行列 H と行列 U を式 (4) と定義して，2 つの行列の積 HU [30]を式 (5) に示して，説明する．式 (4) における $y_{i,j}$ は，$y_{i,j} = h_{i,1}u_{1,j} + \cdots + h_{i,M}u_{M,j}$ である．また，式 (4) における基底数は M に設定した．

$$H = \begin{bmatrix} h_{1,1} & \cdots & h_{1,M} \\ \vdots & \ddots & \vdots \\ h_{N,1} & \cdots & h_{N,M} \end{bmatrix}, U = \begin{bmatrix} u_{1,1} & \cdots & u_{1,K} \\ \vdots & \ddots & \vdots \\ u_{M,l} & \cdots & u_{M,K} \end{bmatrix} \tag{4}$$

$$HU = \begin{bmatrix} y_{1,1} & \cdots & y_{1,K} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,K} \end{bmatrix} \tag{5}$$

NMF アルゴリズムの場合，観測ベクトルを並べた行列である観測行列 Y を変換特徴量となる基底 H と各基底 H に対する重み付けの行列である係数行列 U の積に行列分解する．ここで，行列分解の解は，一意的でなく，一般に誤差が発生する．したがって，NMF における行列分解では，観測ベクトルを並べた行列を規準の定義に応じて H および U の誤差を最小化する行列 H,U を求める最適化問題に帰着すると言える．簡略化した分解表現は，式 (6) で表すことができる．

$$Y \simeq HU \tag{6}$$

28：本節では，作成した観測行列 Y に対する NMF への適用を記載する．他の特徴変換手法への適応は，作成した観測行列 Y を各手法への入力とみなすことで容易に適応できる．詳細は各論文を参照頂きたい [21] [22] [23] [24] [25].

29：(参考) 本研究の範囲対象外だが，数値行列データに変換された行列に対して，平易かつ一意的に機械学習を適用できるツールも存在する．必要に応じて参照されたい．
WEKA - The workbench for machine learning：
https://www.cs.waikato.ac.nz/ml/weka/

30：式に示す行列式は，参考である．分解対象の入力行列 Y は，転置行列である場合もある．

ところで，行列 H,U の誤差を最小化する最適解が所望の解であるためには，背後にある観測行列 Y の生成プロセスにあった適切な規準である必要がある．本研究では，システムが出力するログが対象である．したがって，観測行列の生成プロセスは明らかになっていない．NMF における乖離度規準には，最小二乗誤差 (Frobenius)，一般化 Kullback-Leibler ダイバージェンス，Itakura-Saito ダイバージェンスなどがある．本研究における乖離度規準には，最も一般的な最小二乗誤差を用いた．

4 実　　装

4.1 実験環境

本研究の実装は，プログラミング言語 Python を用いた．ダミー変数への変換処理および項目マッピング処理して観測行列 Y を作成する処理の実装には，Pandas[31]を用いた．本研究におけるグラフ描画には Matplotlib[32]を用いている．ここで，評価を行う場合，作成された観測行列 Y における列は，ダミー変数の数値列と値を直接用いる数値列が混在した観測行列 Y を想定しなくてはならない．本研究では，数値のアプリケーション ID やポート番号などは，値を直接用いずにダミー変数化している．ダミー変数の場合は，値は数値であり頻度であるゆえに，数値の値は 0 から 1 の値に正規化する処理を行った．正規化処理の実装，評価における各アルゴリズムの実装は，scikit-learn[33] [34] を用いた．

4.2 データ・セット

本節では，本研究で用いる IDS のデータダンプに関して説明する．IDS のデータダンプは，2017 年 1 月 1 日から 2017 年 10 月 31 日までのデータセットであり，アラートの総数は 564,561 件から構成される単一のアプライアンスのログデータである．評価実験で用いた IDS のデータダンプにおける項目を表 3 に示す．

表 3　IDS のダンプデータ)

項番	項目名	データ型	項目数
1	over all Apps 項目	**int/str**	11
2	CEF Apps 項目	**str/int/float**	33
3	正解判別用項目	**str**	1
4	アプライアンス項目	**str/int/float**	3

表 3 のダンプデータのファイル形式は *json* ファイルで提供されている．本研究では，*json* ファイルに対して解析処理を実施し，*csv* ファイル化した後，提案手法を用いて，ダミー変数化と項目マッピング処理を実施した．また，提案手法では，項目マッピング処理を行うことで，値のユニーク数が多い項目は処理から除外される．しかし，既存研究における IP アドレス項目などは，値のユニーク数が多い場合であっても，IP レピュテーションなど値と異なる要素で，評価されることもあることから重要度が高い判断要素として使用される場合がある．

ゆえに，本研究における評価では，項目マッピング処理を行い作成した観測行列 Y に加えて，一部値のユニーク数が多い項目を使用した観測行列 Y を作成して，提案手法の比較実験を行った．観測行列 Y だが，提案手法で作成した観測行列 Y は，564,561 行 1,435 列である．本研究では，比較実験に値のユニーク数が多い項目を追加した場合の観測行列 Y は，564,561 行 2,256 列である．本研究では，追加項目として，ソース IP アドレスおよび変換日付の項目を追加した．分類結果の比較を行うことで，追加項目の影響を考慮できる．本研究における評価では，2017 年 1 月 3 日から 2017 年 10 月 25 日までに発生した 845 件のアラートから構成されるデータ集合を検出すべきアラートとした．そして，検出すべきアラートと IDS のダンプデータの共通項目で，マッチング処理を実施して，共通項目が一致した場合に，正解アラートデータとして，正解ラベルを付与した．正解ラベルの一致件数は 436 件である．

4.3 評価方法

本研究では，不要なアラートのスクリーニングが主要なタスクである．アラートスクリーニングにおける正解アラートデータと，アラートデータを比較して，機械学習手法を用いて，正解アラートデータが明らかとなることで，目的が達成できる．また，平易かつ一意的に実現することで，他分野へのデータ解析タスクへの活用も期待できる [35] ．ここで，一般にシステムにおけるアラートの場合，発生するアラートの種類は 1 種類とは限らず，外部からの攻撃起因のアラートやハードウェアが起因したアラートなど，アラートにおけるインシデントの種別に応じて多クラスに分類されることも十分に想定される．そこで，まず，本研究における評価では，提案手法である観測行列 Y の妥当性確認とした．ゆえに，本研究における評価では，特徴量変換を主要タスクとする．特徴量変換を主要タスクとすることで，提案手法で作成した行列の妥当性と提案手法における課題を明らかにすることが期待できる．また，変換特徴量のプロットから，特徴量変換の有効性や特徴量の分布が考察することも期待できる．加えて，特徴量自体の重要性評価と，分類アルゴリズムの有効性も考察した．本研究で評価に用いる機械学習アルゴリズムを表 4 に示す．

表 4　本研究における使用アルゴリズム一覧

項番	アルゴリズム名および参考文献
1	ランダム・フォレスト (RF : Random Forest) [27]
2	バギング (Bagging) [28]
3-1	非負値行列因子分解 (NMF : Non-Negative Matrix Factorization) [20]
3-2	主成分分析 (PCA : Principal component analysis) [21]
3-3	因子分析 (FA : Factor Analysis) [22]
3-4	独立成分分析 (ICA : Independent Component Analysis) [23]
3-5	ランダム射影 (RP : Random Projection) [24]
3-6	t 分布型確率的近傍埋め込み法 (tSNE : t-distributed Stochastic Neighbor Embedding) [25]

各特徴量の重要性の評価では，表 4 における項番 1 のランダム・フォレストの変数の重み付けの値を用いて特徴選択した．

31 : pandas : https://pandas.pydata.org/

32 : Matplotlib : https://matplotlib.org/

33 : scikit-learn : https://scikit-learn.org/stable/

34 : scikit-learn - Official source code repo : https://github.com/scikit-learn/scikit-learn

35 : 価値観や社会構造が変化している中での，課題解決に向けて，多くの取り組みも行われてきている．長期的な計画の実現に向け，異なる分野では，既存データにおける活用タスクなど多様な検討が行われている．
(参考) 国土交通省 国土計画局／ 2030 年の日本のあり方 : http://www.mlit.go.jp/kokudokeikaku/futurevision/

本研究における分類評価では，観測行列の妥当性を明らかにするために，項番 2 のバギングを用いた．特徴量変換では，多様な種類の特徴変換方法が提案されているが，評価対象とするデータの生成プロセスや，成分分析を行う分析対象に応じて，最適なアルゴリズムは異なる．本研究では，表 4 における項番 3-1 から項番 3-6 の 6 種類のアルゴリズムを用いた [36]．評価においては，基底数 M を固定値 2 に設定することで，結果を 2 次元で図示した．成分を図示することで，観測行列に対するアルゴリズムの効果を明らかにすることができる．

4.4 本研究の評価における貢献

本研究では，出力されたシステムのログを数値型の多次元データに変換して，変換特徴量である成分を評価して，明らかにすることが提案の目的である．近年における国際博覧会 (通称:万博) [37] [38] などでは，超スマート社会を目標としており，そうした場では，新しい技術や商品が広がるきっかけとなるとされている．各企業において必ず出力されるシステムのログデータは，可視化しづらいにも関わらず，頻繁に出力される．そうした場で企業で得られたデータを活用するためには，商品などは提案がある程度理解できる状態になっている必要がある．ゆえに可視化は重要である．例えば，インターネット接続デバイスを検索して，可視化する検索エンジンなどがある [39]．新製品の認知媒体などは，1995 年以降の動きでは,「テレビ」は変わらず高位で安定しているが,「新聞」「雑誌」などの紙媒体は低下傾向にある [29]．したがって，情報化が進み，情報アクセス機器が変化してきていることで，可視化などの表現方法で，認知に差が出てきていることがわかる．

近年では，商品の認知媒体にソーシャルメディアなども用いられている．共感に基づいた消費者生活行動モデルに SIPS(Sympathize Identify Paricipate Share & Spread) といった概念も提唱されている [30] [40]．データ分析など情報通信社会のための共通の課題解決には，SDGs[41]達成と Society5.0[42]において，多様な取り組みが行われている．一方で，産業分野における特有の業務フローが複雑であり明確化しづらいないなど，産業分野によっては課題が多い分野もある．世界的には，中長期的な企業価値を考慮する ESG 投資が拡大されている．ESG では，環境への取り組み (Environment) 社会的課題への取り組み (Social)，企業統治への対応，コンプライアンス，情報公開，社会取締役の選任 (Govenance) が企業への課題として挙げられている．産業と技術革新の基盤を作るための目標も掲げている ESG は，重要視されている．2016 年における世界の ESG 市場の規模は 23 兆ドル (2700 兆円) であり，市場規模は大きい．

なお，SDGs 達成と Society5.0 においては，生物多様性 [43]，観光・地域活性化 [44] などの施策や生物多様性における絶滅危惧種保護の施策 [45]，多様な面での不平等をなくそう [46] など産業・技術などと異なる分野における多様な活動も行われている．加えて，トップレベルマネジメント向けの活動 [47] および認知度向上の取り組み [48] も行われている．実際の現場におけるデータ活用では，ログデータの場合は，可視化するなどマーケティング担当者や広告担当者がまず，データを解釈できる状態にする必要がある．ゆえに，本研究で主要タスクとしたログデータから成分分析を行い，可視化するために必要な手法などは，今後の発展における，重要な貢献の一つである．

5 評価実験

5.1 特徴選択と分類評価

まず，提案手法を適用して得られた観測行列 Y に対して，特徴量選択アルゴリズムを適用した結果を図 1 および図 2 に示す．

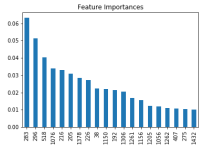

図 1 Random Forest (非ユニーク項目)

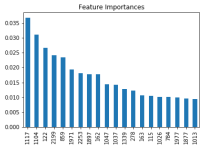

図 2 Random Forest (ユニーク項目使用)

結果から，特徴量選択では，スコアの分布に大きな影響は起きていない．重要な特徴量は，5.2 節および 5.3 節にて，後述する．分類結果だが，表 5 に示す．分類器は，75%のデータでモデル作成，25%のデータで評価した．結果は，Precision 値から，一定の結果が得られた．しかし，Recall 値が十分でないことから，取りこぼしなくアラートを発見することが目的であれば，非ユニーク項目の場合，取りこぼした数は多い．ゆえに，分類における課題は追加項目であることが明らかとなった．

表 5 IDS DataSet Classifier Result

観測行列 Y	Algorithm	Precision	Recall	F-measure
非ユニーク項目	Bagging	0.80	0.63	0.70
ユニーク項目使用	Bagging	0.84	0.70	0.76

そして，本研究における主要タスクである特徴量変換だが，5.2 節および 5.3 節にて，観測行列 Y に対して，特徴量変換を行い，2 次元に図示した結果を示す．観測行列 Y の比較用に特徴量選択された選択特徴量のうち，重み付け上位 10 個の特徴量を表 6 および表 7 に示している．5.2 節の表 3 から表 14 が各手法で 2 次元に変換した変換特徴量の結果である．5.3 節でも，表 15 から表 26 が各手法を用いて特徴量変換した結果である．

36：多くの場合，NMF や PCA などが用いられる．しかし，NMF や PCA などで成分を明らかにできない行列もある．ゆえに，ICA や tSNE などがある．

37： 万博には一般的・総合的な大規模開催の登録博 (最長 6 カ月)，登録博と登録博の間に開催する特定的・専門的な小規模開催の認定博 (最長 3 カ月) がある．

38：(参考) オリンピックの場合は，立候補者は都市，206 の国と地域．万博の場合は，立候補者は国，170 の国と地域．2025 年の大阪・関西万博の来場者想定数は約 2800 万人．経済波及効果約 2 兆円．コンセプトは，人類共通の課題解決に向けた創造・発信．めざすものは SDGs 達成と Society5.0 の実現．

39：Shodan： https://www.shodan.io/

40：SIPS：https://www.dentsu.co.jp/sips/

41：持続可能な開発目標として挙げられている 17 の目標．

42：サイバー空間とフィジカル空間を高度に融合させて，社会的な課題解決を実現することを目標としている．

43：環境省 - 生物多様性民間参画ガイドライン第 2 版： https://www.env.go.jp/nature/biodic/gl_participation/download.html

44：夕張夫妻： https://kawaii.hokkaido.jp/character/yuubarifusai/

45：JWCS トラ保護基金 (現 トラ・ゾウ保護基金 JTEF)： https://www.jtef.jp/

46：LGBT - United Nations Sustainable Development： https://www.un.org/sustainabledevelopment/blog/tag/lgbt/

47：持続可能な開発目標 CEO 向けガイド - WBCSD: https://docs.wbcsd.org/2017/03/CEO_Guide_to_the_SDGs/Japanese.pdf

48：京都国際映画祭 2018： https://2018.kiff.kyoto.jp/news/detail/111

5.2 特徴量変換 - 非ユニーク項目

6 種類の特徴量変換手法を用いて 2 次元の図に特徴量変換した．左側の図が特徴量選択を行わずに特徴量変換した結果，右側の図が特徴量選択を行った後，特徴量変換した結果である [49]．

表 6　特徴選択 (非ユニーク項目)

項番	特徴量番号	特徴量名	重み付け値
1	283	Ccat_fake-av/generic fakeav	0.063302143
2	296	Ccnt_1	0.051229175
3	518	Ccnt_3	0.040428951
4	1076	Cdhost_cdnrep.*******.com	0.033805156
5	216	Csi_file-download	0.03301613
6	205	a_suspicious file download	0.030725006
7	1378	A13L_impact_65	0.028394081
8	226	Ccat_adware/installcore	0.027186445
9	38	Dp_3128	0.02220744
10	1150	Cdhost_rp.*******.com	0.022027154

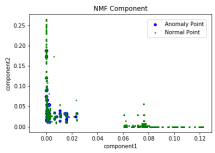

図 3　NMF(特徴選択前)

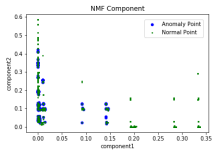

図 4　NMF(特徴選択後)

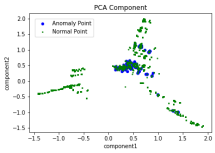

図 5　PCA(特徴選択前)

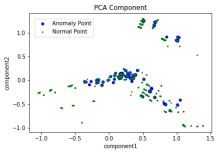

図 6　PCA(特徴選択後)

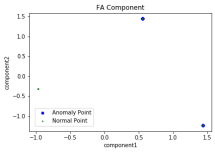

図 7　FA(特徴選択前)

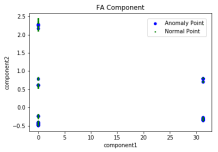

図 8　FA(特徴選択後)

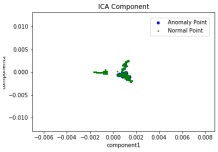

図 9　ICA(特徴選択前)

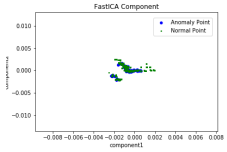

図 10　ICA(特徴選択後)

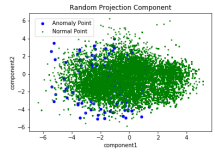

図 11　RP(特徴選択前)

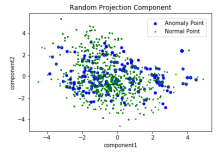

図 12　RP(特徴選択後)

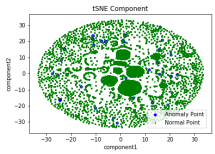

図 13　tSNE(特徴選択前)

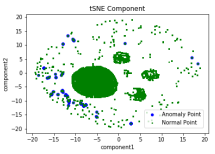

図 14　tSNE(特徴選択後)

49 : 特徴量名が C から開始するのは表 3 の CEF Apps 項目，A13L から開始するのはアプライアンス項目，その他の項目が over all Apps 項目である．

5.3 特徴量変換 - ユニーク項目使用

観測行列 Y に，ユニーク項目を用いている．表 7 から，表 6 にない IP アドレスが重み付けの高い特徴量と評価されていることがわかる [50]．特徴量変換の図の配置は節 5.2 と同じ配置．

表 7　特徴選択 (ユニーク項目使用)

項番	特徴量番号	特徴量名	重み付け値
1	1117	Ccnt_1	0.036794175
2	1104	Ccat_fake-av/generic fakeav	0.031031716
3	122	S_***.***.127.173	0.026577496
4	2199	A13L_impact_65	0.024230831
5	859	Dp_3128	0.02338354
6	1971	Cdhost_rp.*******.com	0.019296312
7	2253	Clv13_6	0.018203253
8	1897	Cdhost_cdnrep.*******.com	0.017843127
9	162	S_***.***.18.20	0.017747201
10	1047	Ccat_adware/installcore	0.01432284

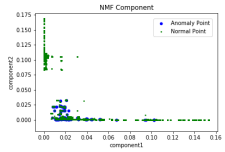

図 15　NMF(特徴選択前)

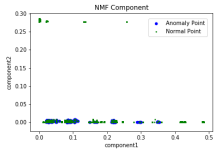

図 16　NMF(特徴選択後)

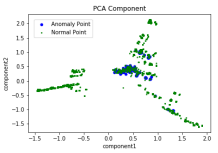

図 17　PCA(特徴選択前)

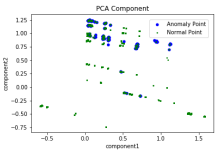

図 18　PCA(特徴選択後)

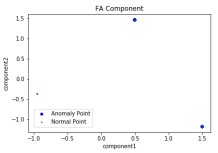

図 19　FA(特徴選択前)

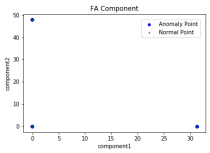

図 20　FA(特徴選択後)

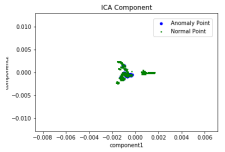

図 21　ICA(特徴選択前)

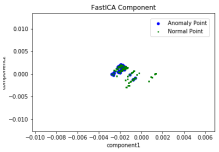

図 22　ICA(特徴選択後)

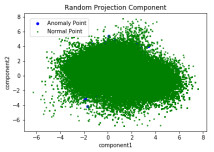

図 23　RP(特徴選択前)

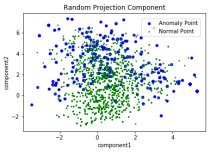

図 24　RP(特徴選択後)

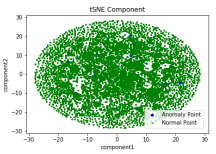

図 25　tSNE(特徴選択前)

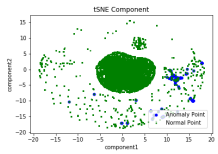

図 26　tSNE(特徴選択後)

50 : IP には，ウォームアップやレピュテーションという概念があり，ISP の重要度判別でも使用されている．(参考)The ABCs of ISPs - SendGrid:
https://ahoy.sendgrid.com/rs/294-TKB-300/images/ABCs_of_ISPs_Guide.pdf

5.4 考　　察

特徴量変換だが，各アルゴリズムで結果が異なることが明らかになった．結果から，クラス分けの場合は，明確にできるのは，NMF か PCA が妥当であり，異常値検出を目的とする場合は，tSNE が効果的であることが明らかになった．一方で，分類評価における，Precision や F-measure では，十分な結果が得られていない．したがって，ダミー変数化後の項目マッピング処理や，項目加工が提案手法における今後の課題である．

6 ま と め

本研究では，IDS のログデータを効率的に行列処理行う手法を提案した．提案手法では，対象となるデータを，ダミー変数化を用いることで，平易かつ一意的にログデータを数値型の観測行列に変換して，評価対象とする手法を提案した．結果，主要タスクとなる特徴量変換による可視化で，ログデータに有効なアルゴリズムを明らかにした．課題としては，項目マッピング処理における最適化，追加項目である．追加項目で分類における精度向上を行いたい．また，提案手法で使用したコーパスは，実際の電子メールのデータセットであるから，既存研究における解析方法を用いた考察を追加したい [51]．また，セキュリティ分野には，セキュリティ特有のデータ検索方法 [52] があることから，分野特有の分析も行いたい．加えて，ログ解析においては，センシングデータを定量的分析する．センシング情報学と呼ばれる分野がある [31]．異なる分野のログデータにも，提案手法を適用して，有効性を検証したい．

謝　　辞

本研究は内閣府の官民研究開発投資拡大プログラム (PRISM) により実施されたものである．ここに謝意を表する．本研究では，提案手法における説明で，カーネギーメロン大学が提供している Enron の電子メールデータセットを利用した．ここに記して，カーネギーメロン大学に感謝申し上げます．

Acknowledgment

This study was supported by the PRISM program of Japan's Cabinet Office (CAO). The author gratefully acknowledge the provide of Enron Email Dataset [53].

文　　献

[1] A. Abduvaliyev, A. K. Pathan, J. Zhou, R. Roman, and W. Wong. On the vital areas of intrusion detection systems in wireless sensor networks. *IEEE Communications Surveys Tutorials*, Vol. 15, No. 3, pp. 1223–1237, Third 2013.

[2] 西岡大, 藤原康宏, 村山優子. 専門知識のないユーザを対象とした情報セキュリティ技術に関する安心感の調査. 情報処理学会論文誌, Vol. 53, No. 9, pp. 2213–2224, sep 2012.

[3] 長山智則. 大規模センサ情報統合に基づく路面・橋梁スクリーニング技術の開発. 計測と制御, Vol. 55, No. 2, pp. 138–144, 2016.

[4] 横谷哲也. Iot と通信ネットワーク技術. 電子情報通信学会誌, Vol. 102, No. 5, pp. 383–387, Dec 2019.

[5] 鷲尾隆. ビッグデータからのモデリング. システム／制御／情報, Vol. 58, No. 1, pp. 3–8, 2014.

[6] 稲見昌彦. 編集長就任にあたって 情報処理 x. 会誌「情報処理」, Vol. 59, No. 4, pp. 316–317, Apr 2018.

[7] 亀井明宏. 電通広告辞典. 電通, 2008.

[8] 児玉紘幸. データマイニングを活用したモノづくりの意思決定支援. 精密工学会誌, Vol. 83, No. 11, pp. 1014–1017, 2017.

[9] 松園和久. 機械学習と iot ストリーミング・データ分析の活用. 日本画像学会誌, Vol. 56, No. 2, pp. 187–191, 2017.

[10] F. Valeur, G. Vigna, C. Kruegel, and R. A. Kemmerer. Comprehensive approach to intrusion detection alert correlation. *IEEE Transactions on Dependable and Secure Computing*, Vol. 1, No. 3, pp. 146–169, July 2004.

[11] L. Spitzner. The honeynet project: trapping the hackers. *IEEE Security Privacy*, Vol. 1, No. 2, pp. 15–23, March 2003.

[12] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita. Network anomaly detection: Methods, systems and tools. *IEEE Communications Surveys Tutorials*, Vol. 16, No. 1, pp. 303–336, First 2014.

[13] Kevin A. Roundy, Acar Tamersoy, Michael Spertus, Michael Hart, Daniel Kats, Matteo Dell'Amico, and Robert Scott. Smoke detector: Cross-product intrusion detection with weak indicators. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017, pp. 200–211, New York, NY, USA, 2017. ACM.

[14] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, Vol. 41, No. 3, July 2009.

[15] 森信介, 笹田鉄郎, Neubig Graham. 確率的タグ付与コーパスからの言語モデル構築. 自然言語処理, Vol. 18, No. 2, pp. 71–87, 2011.

[16] 岩田具治. 機械学習プロフェッショナルシリーズ -トピックモデル. 講談社, 2015.

[17] 竹中平蔵. 経済学から見た電子情報技術者への期待——インベンションからイノベーションへ——. 電子情報通信学会誌, Vol. 100, No. 11, pp. 1156–1159, oct 2017.

[18] 西和彦. 日本の未来のためにいかにして研究開発を活性化するか. 会誌「情報処理」, Vol. 59, No. 3, pp. 220–221, Feb 2018.

[19] 石川周三. 師匠は広告の鬼 - もうひとつの吉田学校. 株式会社宣伝会議, 2007.

[20] 亀岡弘和. 非負値行列因子分解. 計測と制御, Vol. 51, No. 9, pp. 835–844, sep, 2012.

[21] 酒井英昭. 主成分分析と独立成分分析. システム／制御／情報, Vol. 43, No. 4, pp. 188–195, 1999.

[22] 岡田敏彦, 富田真吾. 因子分析法による特徴抽出. 情報処理学会論文誌, Vol. 20, No. 5, pp. 435–443, sep 1979.

[23] 甘利俊一. 情報幾何とその応用-ix : 独立成分分析. システム／制御／情報, Vol. 49, No. 9, pp. 381–386, 2005.

[24] 酒井智弥, 井宮淳. スペクトラルクラスタリングのランダム算法と画像・動画像分割への応用. 電子情報通信学会論文誌 D, Vol. 93, No. 8, pp. 1256–1266, 2010.

[25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9, No. Nov, pp. 2579–2605, 2008.

[26] 亀岡弘和. 非負値行列因子分解の音響信号処理への応用 (小特集近年の音響信号処理における数理科学の進展). 日本音響学会誌, Vol. 68, No. 11, pp. 559–565, 2012.

[27] 馬場真哉, 松石隆. ランダムフォレストを用いたサンマ来遊量の予測. 日本水産学会誌, Vol. 81, No. 1, pp. 2–9, 2015.

[28] Leo Breiman. Pasting small votes for classification in large databases and on-line. *Machine learning*, Vol. 36, No. 1-2, pp. 85–103, 1999.

[29] 株式会社ジェーディーエス. Jnn データバンク - jnn 生活意識レポート 2015 〜全国男女の意識・購買行動からメディア接触まで 88 年以降の時系列変化 〜ライフスタイルトレンドレポート (1988〜2014 年) - 新製品の認知媒体 (時系列/13〜59 歳).

[30] 佐藤尚之, 金田育子, 京井良彦, 信澤宏至, 茂呂譲治, 橋口幸生, 宮林隆吉. Sips 〜来るべきソーシャルメディア時代の新しい生活者消費行動モデル概念〜. 電通モダン・コミュニケーション・ラボ.

[31] 出口光一郎. センシング情報学の構築. 横幹, Vol. 1, No. 2, pp. 80–87, 2007.

51 : Enron Corpus : https://en.wikipedia.org/wiki/Enron_Corpus

52 : Censys : https://censys.io/

53 : The majors that provides Dataset is School of Computer Science. The authors are grateful to Researcher for providing the Enron Email Dataset treated in this paper by The Carnegie Mellon University.