フィルタリング法に基づいたネットワーク異常検知のための特徴選択手法に関する調査

YANG LYU [a)]　YAOKAI FENG [b)]　KOUICHI SAKURAI [c)]

概要：本稿では、特徴選択技術の中で重要かつ広く使われているフィルタリング法に基づく特徴選択技術に焦点を当てる。一般的な説明だけでなく、よく使われる探索アルゴリズムや関連性尺度も詳細に説明する。本稿では、関連するフィルタリング法に基づく特徴選択アルゴリズムを基本的な相関計算法から複雑なアルゴリズムのロジックまでの発展の流れに沿って整理している。

キーワード：ネットワーク攻撃の検出、特徴選択、フィルターテクニック

A Survey on Feature Selection Techniques Based on Filtering Methods for Network Anomaly Detection

YANG LYU [a)]　YAOKAI FENG [b)]　KOUICHI SAKURAI [c)]

Abstract: This paper focuses on feature selection techniques based on filtering methods, which are important and widely used among feature selection techniques. Filtering methods are described not only in general terms, but also in detail with commonly used search algorithms and relevance measures. In this paper, feature selection algorithms based on related filtering methods are organized along the lines of development from basic correlation computation methods to complex algorithmic logic.

Keywords: Network attack detection, Feature selection, Filter Techniques

1. はじめに

現在、ネットワーク攻撃が大きな問題となっていて、組織や個人に巨大な被害や損失をもたらすため、ネットワーク攻撃検知技術は重要な役割を果たしている。特徴選択は、多くの攻撃検知システムで重要なステップであり、不要な特徴の削除により、トレーニングコストを削減し、分類モデルをより理解しやすくし、検知性能を向上させることができる。さらに、特徴選択はモデルサイズを縮小し、検知システムを軽量化することができる。これは、従来の攻撃検知システムを実装できない多数のパワー制限とリソース制限されたデバイスが存在する IoT ネットワークで特に有益である。特徴選択の重要性から、多くの関連技術が提案されており、各技術にはそれぞれの利点と欠点がある。どの技術を使用すべきかは、多くのシステム開発者にとって困難な問題である。特徴選択手法に関する調査論文は多いものの、ほとんどが全体像を捉えようとしていて一般的すぎて、読者に手法の具体的なイメージを把握することが困難である。本論文では、フィルタリング法に基づく特徴選択技術を重点的に扱い、これは現存の特徴選択アルゴリズムの重要な一種で、広く使用されている。一般的な説明に加えて、これらの技術でよく使われる探索アルゴリズと関連性尺度も詳細に説明されている。本論文では、1関連計算方法から特徴選択手法を含む複雑なアルゴリズムのロジックまで、関連するフィルタリング法に基づく特徴選択アルゴリズムを発展の軌跡に沿って整理されている。

2. 特徴選択アルゴリズム

前のセクションで述べたように、特徴選択にはデータ収集と分類モデルのトレーニングのコスト削減、モデルサイズの縮小、分類性能の改善、おそらく分類モデルの理解がよくなるといった多くの利点がある。そのため、多くの特徴選択アルゴリズムが提案されている。

多くの作品では、特徴選択手法はフィルタ技法、ラッパー技法、埋め込み技法 [1][2][3][4][5][6][7][8][9][10][11]. に分類されることがある。しかし、ハイブリッドメソッドやこれらの 3 つのカテゴリに属さない他の手法もある。このセクションでは、それらの簡潔なまとめが提供される。本論文は

a)九州大学大学院システム情報科学府
九州大学マス・フォアイノベーション連携学府
amoyang98@gmail.com@gmail.com

b)九州大学大学院システム情報科学研究院　fengyk@ait.kyushu-u.ac.jp
c)九州大学大学院システム情報科学研究院　sakurai@inf.kyushu-u.ac.jp

フィルタ技法に焦点を当てるが、多くの研究者や開発者にとって多くの利点があり、多くの人が注目するものである。このセクションでフィルタ技法に関する詳細な説明がなされた後、3 つ目と 4 つ目のセクションで、フィルタ技法に不可欠な技術的なコンポーネントである検索アルゴリズムと関連性尺度が説明される。

2.1 フィルタテクニック

フィルタテクニックは、データ固有の性質に基づいて特徴の関連性を評価する。各特徴に対して関連スコアが計算され、スコアが低い特徴が削除される。結果のサブセットの特徴は検出器/分類器への入力として提供される。したがって、特徴の選択のプロセスは、分類アルゴリズムと完全に独立した統計的手法によって最適な特徴のサブセットを決定することである。フィルタテクニックの利点は次のとおりである。

A) データ自身の固有の性質に基づいて特徴の関連性を評価するため、分類アルゴリズムに独立している。

B) 計算効率が高いである。

C) 多くの特徴を持つデータセット (高次元データセット) に簡単にスケールすることができる。

D) 特徴の選択のプロセスは一度だけ行われ、特徴の選択の結果は異なる分類器に使用することができる[4]。

E) ラベル付きトレーニングデータの利用可能性に応じて、監視または非監視の方法で使用することができる。この柔軟性は、広範な IDS アプリケーションに使用することができる[12]。

このように、フィルタ技術は多くの研究者やシステム開発者に魅力的であり、複数の特定の方法が提案されており、それらは 2 つのタイプに分類することができる: 単変量フィルタと多変量フィルタ。前者の方法は個別に特徴を評価し、特徴間の依存関係や相互作用を無視する。したがって、それらは良い特徴選択の結果をもたらすことができない場合がある [13][14]。多変量の方法は特徴間の依存関係や相互作用をいくらか考慮する [2][3][4]。すなわち、フィルタ方法は個別の特徴 (単変量) をランク付けするか、全体の特徴のサブセットを評価することができる (多変量)。フィルタ技法で一般的に使用される評価方法はセクション 4 で詳細に説明される。多変量フィルタ技法の特徴サブセットの生成は、探索戦略に依存する。一般的に、特徴サブセットの生成のための探索戦略には 4 つのタイプがあります: 1) 前向き選択、2) 後方排除、3) 双方向選択、4) ヒューリスティック [15]。フィルタ技法で一般的に使用される探索アルゴリズムは次のセクションで説明される。

サブセットの価値を評価するときには学習アルゴリズムではなく、独立アルゴリズムや統計的測定値を使用する[16]。

単変量フィルタ技術では、各特徴が個別に単変量統計的測定値を使用して重み付けされ、ランク付けされます。これはフィルタベースの特徴ランキングと呼ばれる。一方、多変量フィルタ技術では、複数の特徴をグループとして同時に評価するために多変量の測定値が使用される。これはフィルタベースのサブセット評価と呼ばれる。この場合、探索アルゴリズムを使用して異なるサブセットの価値を比較することができる。

2.2 ラッパーテクニック

ラッパー手法は分類器依存のものであり、分類器との相互作用がある。これらの手法は、特徴サブセットを分類アルゴリズムのパフォーマンスを使って評価する。つまり、各特徴サブセットの評価者は予め定義された分類器（例えば、SVM、Naive Bayes、Random Forest など）である。評価プロセスは各サブセットに対して繰り返され、特徴サブセットの生成はフィルタ手法と同様の探索戦略に依存する。つまり、ラッパー特徴選択は分類アルゴリズムを特徴選択プロセスに埋め込み、すべての特徴のさまざまなサブセットを探索アルゴリズムより生成して評価する。フィルタベースの手法との違いは、分類器が特徴サブセットの評価に使われていることである。これらの手法の普遍的な欠点は、特徴選択プロセスが予め定義された分類モデルと深く相互作用するため、オーバーフィットのリスクが高いこと、評価プロセスが各サブセットに対して繰り返されるため、計算コストが非常に高いことである（大きなデータセットで分類器をトレーニングし、すべての特徴サブセットを評価することが計算コストが高いためです）[4][5]。また、特徴選択の結果は評価された分類アルゴリズムに偏っていることもある。

2.3 埋め込みテクニック

埋め込みテクニックもラッパーテクニックのような分類器に依存しているため、分類器とも相互作用する。しかし、これらの方法では、特徴の最適なサブセットの決定は分類器の構築に組み込まれている。つまり、埋め込み手法は分類アルゴリズムの実行中に特徴選択を行う。特徴選択のプロセスは、通常の機能または拡張機能として分類アルゴリズムに埋め込まれる。したがって、埋め込み手法は特定の学習/分類アルゴリズムに固有である[4][5]。この文脈で使用される一般的な分類アルゴリズムには、一定の種類の決定木、加重ナイーブベイズ、SVM の点数ベクトル[4]、SVM-RFE[17]、カーネルペナルティ SVM[18]などがある。いくつかの決定木アルゴリズムの種類には、CART（分類と回帰木）[19]、C4.5[20]、ランダムフォレスト[21]などがあり、多項式ロジスティック回帰とそのバリエーション[22]などもある。埋め込み手法は特徴のサブセットの価値を評価するために独立統計量と分類アルゴリズムを包括的に使用する[16]。

多くの論文は上述の 3 つの特徴選択手法のみを言及しているが、実際には、いくつかの研究者はハイブリッド技法やこれら 3 つのカテゴリに分類することが困難な他の技法を提案している。次の 2 つのサブセクションで簡単にまとめられる。

2.4 ハイブリッドテクニック

フィルタとラッパーの技法のいくつかのハイブリッドアプローチも提案されている。ここではフィルタとラッパーの良い性質が組み合わせられている [15][23][24][25]。まず、フィルタリング手法を使って複数の候補サブセットを得る。次に、ラッパーは最適な候補を見つけようとする。

2.5 その他のテクニック

上記のカテゴリに含まれないいくつかの方法が提案されている。これらの方法には、ファジー・ランダムフォレストに基づく特徴選択[26]、ハイブリッド遺伝的アルゴリズム[27]、ハイブリッド蟻コロニー最適化[28]、またはハイブリッド重力探索アルゴリズム[29]が含まれる。

さらに、いくつかの特徴選択方法は、フィットエラーを最小化する目的関数に基づく特徴ウェイト付けに基づいている[30][31]。一般的には線形分類器（例えば、SVM）が使用され、分類にほとんどまたは全く貢献しない特徴がペナルティを受ける。

論文[32]では、個々の特徴選択方法からの独立した結果を組み合わせるアンサンブル手法が提案されている。ヒューリスティックベースとグリーディーベースの 2 つのアンサンブル方法が提示され、選択プロセスを自動化しようとしている。ただし、著者が述べたように、探索を停止する終了条件を決定するのは難しいということである。

3. フィルタベースの特徴選択方法における検索アルゴリズム

特徴選択は、多くの機械学習問題では、すべての可能な特徴を使用すると過学習や高計算複雑度が生じるため、最も予測性能が優れている最小のサブセットを選択することが有用である。検索アルゴリズムは、多数の特徴から最適なサブセットを見つけるのに役立つ。したがって、検索アルゴリズムの品質は対応する特徴選択アルゴリズムの性能に大きく影響する。以下は、特徴選択アルゴリズムが検索アルゴリズムの使用を必要とするシナリオである。

A）高次元データセット：特徴の数が非常に多い場合、すべての特徴のサブセットを評価することは計算上非常にコストがかかる。検索アルゴリズムは、サブセットのみを評価することで、効率的に最も重要な特徴を特定するのに役立つ。

B）過学習：モデルに特徴が多すぎる場合、トレーニングデータに適合することができるが、新しいデータでは性能が低下する。検索アルゴリズムは、過学習を防ぐために最小のサブセットの特徴を見つけるのに役立つ。

C）モデルの解釈性を改善：検索アルゴリズムは、最も情報的なサブセットの特徴を特定するのに役立ち、モデルをより解釈可能にすることができる。

D）計算複雑度の縮小：特徴の数が非常に多い場合、トレーニングと評価を行うことも計算上非常にコストがかかる。重要な特徴のサブセットを特定するための検索アルゴリズムの使用により、計算複雑度が縮小することができる。

E）一般化の改善：検索アルゴリズムは、新しいデータに適合するサブセットの特徴を特定するのに役立つ。

多変量フィルタの特徴サブセット生成には、前向選択、後退除去、双方向検索、ヒューリスティック特徴サブセット選択などの検索戦略が使用される。これらの戦略は、それらがどのように開始し、可能な特徴サブセットの空間をどのように探索するかによって異なり、与えられたタスクのための最良の特徴サブセットを見つけることを目的としている。これらの戦略は、それぞれがどのように開始し、可能な特徴サブセットの空間を探索するかにおいて異なるが、与えられたタスクに最適な特徴のサブセットを見つけることが目的である[15]。

これらの検索戦略は、指数、順次、およびランダムに大まかに分類できる[33]。

指数アルゴリズムには、徹底的な探索と分岐と制限が含まれる。これらの方法はすべての特徴のサブセットを評価するが、大きな特徴空間の場合は計算コストが非常に高いことがある。

シーケンシャルアルゴリズムには、前向選択、後退除去、双方向検索が含まれる。これらの方法は一度に 1 つの特徴を追加または削除し、局所的な最小値に陥ってしまう可能性があり、最適な特徴のサブセットを見つけられない場合がある。

ランダムアルゴリズムには、ヒューリスティック特徴サブセット選択、遺伝的アルゴリズム、シミュレーテッドアニーリング、ランダムサーチが含まれる。これらの方法は、探索手順にランダム性を組み込み、局所的な最小値を回避し、特徴空間をより広い範囲探索することができる。これにより、最適な特徴のサブセットを見つける可能性が高くなる。

この論文では、以下の主流の検索アルゴリズムを紹介する。

3.1 貪欲なヒルクライム

この探索戦略は、現在の特徴のサブセットに対する局所的な変更だけを考慮する。一般的に、局所的な変更は、サブセットから特徴を追加または削除することだけである。初期サブセットが空であり、一度に 1 つの特徴だけが追加される場合、前向選択として知られている。逆に、初期サブセット

が完全なセットであり、一度に 1 つの特徴を削除する場合、後退除去として知られている[34][35]。別のアプローチは、ステップワイズ双方向検索と呼ばれている。追加と削除の両方を使用する。現在のサブセットに対するすべての可能な変更の中から、探索アルゴリズムは最善の変更を選択するか、単に現在のサブセットの利点を改善する最初の変更を選択することができる[36]。いずれの場合も、変更が受け入れられた場合は、再度考慮されない。

3.2 ベストファーストサーチ

ベストファーストサーチは、探索パスに沿ってバックトラックすることができる探索方法である。貪欲な山登りと同様、ベストファーストサーチは現在のサブセットに局所的な変更を加えて探索空間を探索する[37]。しかし、貪欲な山登りとは異なり、ベストファーストサーチは現在探索しているパスが望ましくない場合でも、より有望なサブセットにバックトラックしてそこから再度探索することができる。この機能は主に 2 つのリストによって実現される。そのうち、オープンリストは現在の状態のサブステートを記録し、クローズドリストは以前の状態を記録する。図 3 はベストファーストサーチアルゴリズムを示している。

3.3 遺伝的アルゴリズム

遺伝的アルゴリズムは、生物においての選択原理に基づいた適応的な探索アルゴリズムである[38]。アルゴリズムは競合する解の集合から開始し、時間の経過とともに最適解に収束する。探索戦略は解空間における並列探索で、局所的な最適解を避けることができる。次世代の新しいサブセットを生成する各操作には交叉と変異が含まれ、選択メカニズムは新しい個体の適合度に基づいている。適合度が高いほど、選択される確率が高くなる。このプロセスは終了条件が満たされるまで繰り返される。特徴選択では、解は通常、サブセットを表す遺伝的な固定長のバイナリ文字列で、文字列の各位置の値は特定の特徴の存在または不在を示す。

4. フィルタメソッドに対する関連性尺度

上述で、フィルタメソッドにおいて、特徴選択アルゴリズムは（1）フィルタベースの特徴ランキング、（2）フィルタベースのサブセット評価に分類されることを述べた。どちらも、独立したアルゴリズムや統計的な測定に基づいて、特徴または特徴のサブセットの品質を判断する必要がある。そのため、このセクションでは、分類に対する特徴のメリットを評価する方法について議論する。一般的に、良い特徴はクラスと関連しているが、他の特徴とは冗長ではない[36]。この定義に基づいて、特徴選択問題は、特徴の目的変数への関連性を計算するための 適切な方法を見つけ、それに基づいた合理的な特徴選択スキームを探すことに言い換えられる。

現在、特徴間の相関を計算する複数の方法がある。これには、線形メソッド（ピアソン、スピアマン相関係数）、非線形メソッド（ユークリッド、マンハッタン、角度コサイン相関係数）などがある。カイ二乗検定や相互情報も使用できる。方法の選択は特定のシナリオや要件による。

4.1 ピアソン相関

ピアソンの相関係数（PCC）は、2 つのランダム変数の間に存在する線形の関係の深さと方向を測定する統計的手法です[39]。2 つの連続する変数 X と Y について、PCC（X、Y）は以下の式（1）を使って計算されます。

$$PCC(X,Y) = \frac{\sum_i (x_i - \bar{x_i})(y_i - \bar{y_i})}{\sqrt{\sum_i (x_i - \bar{x_i})^2 \sum_i (y_i - \bar{y_i})^2}}, \quad (1)$$

ここで $\bar{x_i}$ は X の平均、$\bar{y_i}$ は Y の平均である。

PCC(X,Y)の値は-1 から 1 までの正規化された範囲になる。値が-1 または 1 の場合、2 つの変数間に強い相関があることを意味し、0 の場合、2 つの変数は相互独立であることを意味する。

PCC は非線形な依存変数の関係を明らかにすることができないため、またすべての特徴量に値がある必要があるため、別の相関測定方法として、以下に述べられたカイ二乗、情報利得（IG）、相互情報（MI）などが提案されている。

4.2 カイ二乗

統計学において、カイ二乗検定は 2 つの確率変数の関係性を決定するための方法である。特に、2 つの変数が独立していると仮定され（帰無仮説）、実際の値と理論的な値（2 つの変数が本当に独立している場合に持つべき値）の誤差を観察する。誤差が十分に小さければ、帰無仮説が受け入れられる。誤差が一定のレベルに達すると、このような誤差は偶然または不正確さによって引き起こされることはないと考えられ、2 つの変数は実際に相関していると考えられる。すなわち、帰無仮説は拒否され、代替仮説が選択される。カイ二乗検定は式（2）で定義されている [40][41]

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}, \quad (2)$$

ここで、

K = number of (no.) classes,

A_{ij} = no. patterns in the ith interval, jth class,

R_i = no. patterns in the ith interval = $\sum_{j=1}^{k} A_{ij}$,

C_j = no. patterns in the jth class = $\sum_{i=1}^{2} A_{ij}$,

N = total no. patterns = $\sum_{i=1}^{2} R_i$,

E_{ij} = expected frequency of A_{ij} = $R_i * C_j / N$

特徴選択において、各特徴と各カテゴリ間のカイ二乗検定を実施するだけでよく、その結果を降順で並べ替える。最後

に、相対的に大きいカイ二乗値を持ついくつかの特徴を選択する。

カイ二乗検定にもそれ自身の欠陥がある。特定のカテゴリのインスタンスに特定の特徴が存在するかどうかを数えるが、特徴がインスタンスに何回出現するかは数えない。これは低頻度の単語に偏ってしまう。したがって、カイ二乗検定は通常他の計算方法（IG など）と組み合わせて強みを最大化し、弱点を避ける必要がある。

4.3 情報利得 (IG)

変数の不確実性を示すのがエントロピーである。情報利得は特定の条件下での変数の不確実性の減少度を示す。式(3)は変数 X のエントロピーを定義し、式(4)は Y の発見が判明したときの X の条件付きエントロピーを定義する。

$$H(X) = -\sum_{x\in X} p(x)\, log_2(p(x)); \quad (3)$$

$$H(X|Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\, log_2(p(x|y)). \quad (4)$$

式 (2) と (3) を組み合わせると、情報ゲインの式 (5) が得られます [42]。

$$IG(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \quad (5)$$

ここで、式 (5) は対称であるため、結果は変数の順序に依存しない。

4.4 相互情報 (MI)

MI（相互情報）[43]は、2 つの変数間の相互依存関係を表す指標で、連続確率変数の場合は式(6)、離散確率変数の場合は式(7)によって定義される。 MI と IG は同じ式を持っていることが見られる。

$$I(X;Y) = H(X) - H(X|Y); \quad (6)$$

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\, log\, \frac{p(x,y)}{p(x)p(y)}. \quad (7)$$

4.5 多変量相互情報量 (MMI)

S. Mohammadi ら [44] は、同時に複数の変数を計算できる相互情報の計算方法を提案した。つまり、式(8)である。

$$I_N(X_1;X_2;\ldots;X_N) = \sum_{k=1}^{N}(-1)^{K-1} \sum_{\substack{X\in(X_1;X_2;\ldots;X_N)\\ |X|=k}} H(X). \quad (8)$$

4.6 相互情報特徴選択 (MIFS)

Battiti の算法（MIFS）[45]は、式（9）で定義されており、$I(C;f_i)$と$I(f_s;f_i)$を計算して関連する特徴を選択するために MI を適用する。

$$J_{MIFS} = I(C;f_i) - \beta \sum_{f_s\in S} I(f_s;f_i), \quad (9)$$

ここで、f_iは元の特徴セットに属し、f_sは選択された特徴部分集合に属する。パラメータ β は冗長性に関連しており、最適な特徴部分集合の選択に重要な影響を与える。ただし、冗長性パラメータ β の適切な値の選択はまだ解決されていない問題である。

4.7 多変量相互情報ベースの特徴選択 (MMIFS)

S. Mohammadi ら [44] は、特徴選択のための提案されたアルゴリズムとして、式 (9) を式 (10) に変更しました。

$$E_{MMIFS} = argmax_{f_i\in F}\big(MI(C;f_i) - \beta * I_N(f_i;f_s)\big), \quad (10)$$

ここで、$I_N(.)$ は前に定義された MMI 関数である。

このアルゴリズムのアイデアは、最初に元のデータセット内の各特徴とカテゴリ間の MI を計算することから始まる。そして、最も高いスコアを持つ特徴を選択されたサブセットに入れる。その後、式（10）を通じて、残りの特徴から最も高い値を持つ特徴を選択し、目的の数に達するまで選択されたサブセットに追加する。

4.8 相関ベースの特徴選択 (CFS)

CFS のコアは、次の前提に基づくヒューリスティックな方法で特徴サブセットの価値を評価することである:最適な特徴サブセットにはクラスと高い相関関係を持つが相互に相関関係を持たない特徴が含まれている。評価基準は式（11）[37]で示される。

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}, \quad (11)$$

ここで、M_S は k 個の特徴を含む特徴サブセット S の「メリット」、$\overline{r_{cf}}$ は特徴クラスの相関の平均、$\overline{r_{ff}}$ は特徴と特徴の相互相関の平均である。 ここで、 $\overline{r_{xy}}$ は、ピアソンの相関係数や対称不確実性 (Symmetric Uncertainty) [46] などで、相関を測定して得るメトリックである。対称不確実性は式(12)で示される。

$$SU(X,Y) = 2\left[\frac{H(X)-H(X|Y)}{H(X)+H(Y)}\right]. \quad (12)$$

この式の分子は実際の情報利得であることがわかる。実際、情報利得はより価値のある特徴に偏りがある。したがって、式 8 は分母を通じてこのバイアスを補正し、値を[0,1]の範囲に正規化する。値 1 は 2 つの変数が強く相関していることを示し、値 0 は 2 つの変数が互いに独立していることを示す。

4.9 効率的な相関ベースの特徴選択 (ECOFS)

W. Wang ら[47]は、CFS 内の特徴間のペアワイズ計算を避け、MIFS で適切なβ値を設定する負担を取り除くために、効率的な相関ベースの特徴選択基準を提案した。これは式（13）で定義される。

$$J_{ECOFS} = SU_{f_i \in F}(f_i, c) - \max_{f_s \in S}(SU(f_s, f_i)), \qquad (13)$$

ここで、f_i と f_s は式 (9) と同じように定義される。$\max_{f_s \in S}(SU(f_s, f_i)$ は、特徴 f_i と選択された特徴 f_i の間の測定された冗長性を表す最大値である。彼らが提案したアイデアは、特徴がクラスに対する貢献が選択された特徴間の特徴の冗長性より大きい場合、特徴を「良い」ものとみなし、それを保持するという原則に従う。

5. フィルタベースの特徴選択手法のパフォーマンスに関する比較研究

表 1 は、2016 年以後の論文に含まれているフィルターベースの特徴量選択技術の比較を示している。それ表に含まれる情報は、使用された FS 方法、あるデータセットに対して選択された特徴量の数、使用された検出方法、使用されたデータセット、および技術の性能指標である。技術の性能指標には、検出モデルの正確性率と生成時間が含まれている。

表 1 2016 年以後のフィルターベースの特徴選択手法の比較

Author/Year	FS method	No.of features	Detection method	Dataset	Performance metrics
Shahbaz et al. (2016)[48]	CFS	4	J48	NSL-KDD	ACC(%): 86.1 , Time(sec):<15
Ullah et al (2017)[49]	IG	ISCX: 4 NSL-KDD: 6	J48	ISCX NSL-KDD	ACC(%): 99.70(ISCX) 99.90(NSL-KDD) , Time(sec): 15
Kushwaha et al. (2017)[50]	MI	5	Support vector machine (SVM)	KDDcup'99	ACC(%): 99.91
Moham madi et al. (2018)[44]	MI	/	least square version of SVM (LSSVM)	KDDcup'99, NSL-KDD, Kyoto 2006+	ACC(%): 94.31(KDDcup'99) 98.31(NSL-KDD) 99.11(Kyoto 2006+),
Wang et al. (2019)[46]	Efficient CFS (MIFS + Symmetric Uncertainty)	KDDcup'99: 9 NSL-KDD: 10	One-class SVM	KDDcup'99, NSL-KDD	ACC(%): 99.85(KDDcup'99) 98.64(NSL-KDD), Time(sec): 5.3(KDDcup'99) 1.7(NSL-KDD)

6. まとめ

特徴量選択の重要性を紹介した後、この論文では既存の特徴量選択技術の概要と分類を提供する。その後、重要で広く使用されているフィルターベースの方法について、一般的な説明、検索アルゴリズム、およびそのような方法で一般的に使用される関連性尺度を含めて詳しく説明する。

私たちは、基本的な相関計算方法から相関計算と特徴量選択手順の両方を含む複雑なアルゴリズムの論理まで、関連するフィルターベースの特徴量選択アルゴリズムを開発軌跡に沿って整理した。変数間の相関を測定するために、線形関係に基づくピアソン相関係数および統計に基づくカイ二乗検定の制限のため、最近の論文で最も一般的に使用されるメトリックは、情報理論に基づく情報利得または相互情報量である。

この作業の焦点は、相関に基づいて特徴量を効果的にスクリーニングする方法と、対応するルールを策定することにある。この論文は、研究者やシステム開発者の有用なガイドとして機能し、読者が既存の特徴量選択技術の一般的な概要を持つだけでなく、広く使用されているフィルターベースの特徴量選択技術のより具体的な理解を得ることができるようにする。

謝辞 本研究は、日本学術振興会とインド政府科学技術省・科学技術庁(DST)の日印共同科学プログラム（2022-2024)の支援を受けたものである。

参考文献

[1] Sharma, N., & Arora, B. (2022). A critical review of feature selection techniques for network anomaly detection: Methodologies, challenges, evaluation, and opportunities. https://www.researchsquare.com/article/rs-1940841/v1 (accessed on Jan. 26, 2023)

[2] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-2003) (pp. 856-863).

[3] Senliol, B., Gulgezen, G., Yu, L., & Cataltepe, Z. (2008, October). Fast Correlation Based Filter (FCBF) with a different search strategy. In 2008 23rd international symposium on computer and information sciences (pp. 1-4). IEEE. doi: 10.1109/ISCIS.2008.4717949.

[4] Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

[5] Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. Pertanika Journal of Science & Technology, 26(1).

[6] Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371-6385.

[7] Ladha, L., & Deepa, T. (2011). Feature Selection Methods and Algorithms. International Journal on Computer Science and Engineering (IJCSE), 3(5), 1787 – 1797.

[8] Cantu-Paz, E. (2004). Feature subset selection, class separability, and genetic algorithms. In Genetic and Evolutionary Computation–GECCO 2004: Genetic and Evolutionary Computation Conference, Seattle, WA, USA, June 26-30, 2004. Proceedings, Part I (pp. 959-970). Springer Berlin Heidelberg.

[9] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34, 483-519.

[10] Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Information sciences, 282, 111-135.

[11] Thakkar, A., Lohiya, R. (2022). A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. Artif Intell Rev 55, 453–563. https://doi.org/10.1007/s10462-021-10037-9

[12] Sánchez-Maroño, N., Alonso-Betanzos, A., & Calvo-Estévez, R. M. (2009, June). A wrapper method for feature selection in multiple classes datasets. In International Work-Conference on Artificial Neural Networks (pp. 456-463). Springer, Berlin, Heidelberg.

[13] Piao, Y., Piao, M., Park, K., & Ryu, K. H. (2012). An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics, 28(24), 3306-3315.

[14] Yusta, S. C. (2009). Different metaheuristic strategies to solve the feature selection problem. Pattern Recognition Letters, 30(5), 525-534.

[15] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). IEEE. doi: 10.1109/MIPRO.2015.7160458.

[16] Zuech, R., & Khoshgoftaar, T. M. (2015, August). A survey on feature selection for intrusion detection. In Proceedings of the 21st ISSAT International Conference on Reliability and Quality in Design (pp. 150-155).

[17] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46, 389-422.

[18] Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. Information Sciences, 181(1), 115-128.

[19] Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

[20] Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78.

[21] Sandri, M., & Zuccolotto, P. (2006). Variable selection using random forests. In Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005 (pp. 263-270). Springer Berlin Heidelberg.

[22] Cawley, G., Talbot, N., & Girolami, M. (2006). Sparse multinomial logistic regression via bayesian l1 regularisation. Advances in neural information processing systems, 19.

[23] Das, S. (2001, June). Filters, wrappers and a boosting-based hybrid for feature selection. In Icml (Vol. 1, pp. 74-81).

[24] Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. Expert Systems with Applications, 38(7), 8144-8150.

[25] NAQVI, S. (2011). A Hybrid filter-wrapper approach for FeatureSelection.

[26] Cadenas, J. M., Garrido, M. C., & MartíNez, R. (2013). Feature subset

selection filter–wrapper based on low quality data. Expert systems with applications, 40(16), 6241-6252.

[27] Oh, I. S., Lee, J. S., & Moon, B. R. (2004). Hybrid genetic algorithms for feature selection. IEEE Transactions on pattern analysis and machine intelligence, 26(11), 1424-1437.

[28] Ali, S. I., & Shahzad, W. (2012). A feature subset selection method based on conditional mutual information and ant colony optimization. International Journal of Computer Applications, 60(11), 5-10.

[29] Sarafrazi, S., & Nezamabadi-Pour, H. (2013). Facing the classification of binary problems with a GSA-SVM hybrid system. Mathematical and Computer Modelling, 57(1-2), 270-278.

[30] Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. Briefings in bioinformatics, 9(5), 392-403.

[31] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320.

[32] Nakashima M., Sim A., and et al. (2021). Automated Feature Selection for Anomaly Detection in Network Traffic Data. ACM Transactions on Management Information Systems 12 (3), 1–28, https://doi.org/10.1145/3446636

[33] Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

[34] Kittler, J. (1978). Feature set search algorithms. Pattern recognition and signal processing.

[35] Miller, A. (2002). Subset selection in regression. chapman and hall/CRC.

[36] Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

[37] Winston, P. H. (1984). Artificial intelligence. Addison-Wesley Longman Publishing Co., Inc..

[38] Holland, J. H. (1992). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press.

[39] Teukolsky, S. A., Flannery, B. P., Press, W. H., & Vetterling, W. T. (1992). Numerical recipes in C. SMR, 693(1).

[40] Li, Y., Fang, B. X., Chen, Y., & Guo, L. (2006, November). A lightweight intrusion detection model based on feature selection and maximum entropy model. In 2006 International Conference on Communication Technology (pp. 1-4). IEEE.

[41] Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE international conference on tools with artificial intelligence (pp. 388-391). IEEE.

[42] Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn 16, 235–240 (1994). https://doi.org/10.1007/BF00993309.

[43] Cover, T. M., & J. A. Thomas, Elements of Information Theory. Hoboken, NJ, USA: Wiley, 2012.

[44] Mohammadi, S., Desai, V., & Karimipour, H. (2018, October). Multivariate mutual information-based feature selection for cyber intrusion detection. In 2018 IEEE electrical power and energy Conference (EPEC) (pp. 1-6). IEEE.

[45] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks, 5(4), 537-550.

[46] Teukolsky, S. A., Flannery, B. P., Press, W. H., & Vetterling, W. T. (1992). Numerical recipes in C. SMR, 693(1).

[47] Wang, W., Du, X., & Wang, N. (2018). Building a cloud IDS using an efficient feature selection method and SVM. IEEE Access, 7, 1345-1354.

[48] Shahbaz, M. B., Wang, X., Behnad, A., & Samarabandu, J. (2016, October). On efficiency enhancement of the correlation-based feature selection for intrusion detection systems. In 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 1-7). IEEE.

[49] Ullah, I., & Mahmoud, Q. H. (2017, December). A filter-based feature selection model for anomaly-based intrusion detection systems. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2151-2159). IEEE.

[50] Kushwaha, P., Buckchash, H., & Raman, B. (2017, November). Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 839-844). IEEE.