2007 年 5 月 8 日(火)
1031060322 松本　亮介

# ＳＶＭを用いた
# ネットワークトラフィックからの異常検出（概略）

## 0 はじめに

※この資料は，参考文献を元に，参考文献の概略を作成したものです．この研究を私が実際に行ったわけではありませんので注意してください．

## 1 流れ

従来の異常検出の手法では，計算の処理に時間がかかるものや，対象とするデータを限定しているため誤報率が高いなどの問題があった．これに対して，最近では，すべてのネットワークトラフィック変化を反映する特徴量をもとに，主成分分析を用いる手法などが提案されている．本稿では，パケットのヘッダ情報を属性ごとに数値化してベクトルデータとして表現し，SVM（サポートベクターマシン）を用いてクラスタリングすることにより，異常を検出する方法を提案する．そして，主成分分析を用いた手法とＳＶＭを用いた手法を比較・評価していく．

## 2 手法

まず，主成分分析を用いた手法について簡単に説明する．これは，プロトコルごとのパケット間の相関関係に基づき，主成分軸から距離的に離れたものを異常として検出する手法である．その時に，以下の２つの仮定に基づいている．

1. 正常状態は，ある超平面に沿って分布する
2. 異常状態は，プロトコルの分布が崩れ，超平面から離れる

主成分分析とは，多次元データの次元を圧縮する手法の１つで，もとデータからの２乗誤差を最小にする手法である．つまり，データの分散が最大となる超平面を求めることを意味し，その超平面は正常状態が密に分布するものとなる．この特徴を利用して，主成分分析により特徴空間に配置されたネットワーク状態を表すベクトルデータの主成分超平面を求め，主成分超平面との投影距離の大きいものを異常状態として検出している．

しかしこの手法では，多次元データからの２乗誤差を最小にするという点では最適な特徴空間を構成するが，クラス情報を用いないため，クラスタリングを行う上では必ずしも最適ではない．そこで，クラス分離度が大きくなるような属性を抽出する手法の一つである SVM に着目し，主成分分析のかわりに SVM を用いて識別超平面を導くことを考える．

SVM では，線形分離可能な場合，識別関数は与えられた次元の学習データと各データのクラスラベルで表すことができる．識別面は[識別関数]＝0 で表される超平面で与えられる．そして，超平面と学習データの距離の最小のものをマージンと呼ぶ．この超平面とすべての学習データが正しく分離されるためには，制約条件が必要である．その制約条件を満たす超平面は無数に存在するが，SVM ではマージンを最大にする超平面を識別平面とする．

線形分離が不可能な場合，ある適当な非線形変換によって，もとの次元よりも高次元な空間へ写像することにより，線形分離性を高めることができる．写像した先の高次元空間で線形分離を行えば，もとの空間で非線形分離を行うのと等価になる．そこで，もとの空間で定義されるカーネル関数を用意し，それをある一定の条件で計算すればよい（論文 2.3.3 参照）．このような条件を満たすカーネルには，多項式カーネルや RBF カーネルの２つがある．

## 3 結果

実験では，観測点を通過するパケットのヘッダ情報の種別ごとに単位時間当たりのパケット数をカウントし，ネットワークトラフィックを多次元ベクトルデータとして数値化する．これにより，単位時間におけるネットワークの状態をある次元空間の点として表現することができる．これにより，単位時間でのパケット種別ごとの分布が明らかになり，通信状態を把握することができる．このベクトルデータ

をパケット種別ごとに平均 0, 標準偏差 1 となるように正規化し，主成分分析したものと，SVM を用いて評価したものを比較する．

主成分分析による異常検出では，特徴空間の主成分軸と投影距離をある閾値で検出した時，閾値を大きくすると false positive（異常状態を誤って正常状態と判定）は減少するが，false negative（正常状態を誤って異常状態と判定）増加する．このため，全体の誤り率を小さくするための閾値を設定するのは困難である．これは，単位時間当たりのポートスキャンの数が少ない場合，主成分分析により構成される特徴空間において，単純に主成分軸との投影距離だけでは異常状態が分離できないことが原因と予想できる．

これに対して，SVM を用いた手法では，85%程度の異常検出が可能であり，主成分分析による実験とは異なり false positive はまったく検出されていない．また，false negative と判定されたものを見ると，これも主成分分析と同じように，単位時間あたりのポートスキャン数が少ない場合，正常状態と統計的差異が表れないため，正常状態として判定されていることがわかった．しかし，SVM ではあらかじめデータベースを用意する必要がないという特徴をもっていて，さらに，検出精度が低いとされているポート番号種別情報やパケットサイズ情報を対象とした場合でも，80%程度の異常検出が可能であり，低レイヤの異常検出に関しては，十分な性能であると考えられる．

## 参考文献

[1] 宮本貴朗, “ＳＶＭを用いたネットワークトラフィックからの異常検出”