A Comparison of Feature-Selection Methods for Intrusion Detection

Hai Thanh Nguyen, Slobodan Petrović and Katrin Franke

Gjøvik University College, Norway

Introduction

- The problem of intrusion detection
  - Analyzed as a pattern recognition problem
    - Has to tell normal from abnormal behavior of network traffic and/or command sequences on a host
    - Classifies further abnormal behavior to undertake adequate counter-measures

Introduction

- Models of IDS usually include
  - A representation algorithm
    - Represents incoming data in the space of selected features
  - A classification algorithm
    - Maps the feature vector representation of the incoming data to elements of a certain set of values (e.g. normal, abnormal, etc.)

Introduction

- Some IDS also include a feature selection algorithm
  - Determines the features to be used by the representation algorithm
- If a feature selection algorithm is not included in the IDS model, it is assumed that a feature selection algorithm is run before the intrusion detection process

Introduction

- The feature selection algorithm
  - Determines the most relevant features of the incoming traffic
    - Monitoring of those features ensures reliable detection of abnormal behavior
- The number of selected features heavily influences the effectiveness of the classification algorithm

Introduction

- The task of the feature selection algorithm
  - Minimize the cardinality of selected features without dropping potential indicators of abnormal behavior
- Feature selection for intrusion detection
  - Manual (mostly) – based on expert knowledge
  - Automatic

Introduction

- Automatic feature selection
  - The filter model
    - Considers statistical characteristics of a data set directly
    - No learning algorithm involved
  - The wrapper model
    - Assesses the selected features by evaluating the performance of the classification algorithm

Introduction

- Individual feature evaluation is based on
  - Their relevance to intrusion detection
  - Relationships with other features
    - Such relationships can make certain features redundant
- Relevance and relationship are characterized in terms of
  - Correlation
  - Mutual information

Introduction

- We focus on 2 feature selection measures for the IDS task
  - Correlation feature selection (CFS)
  - Minimal-redundancy-maximal-relevance (mRMR)
- Both feature selection measures contain an objective function, which is maximized over all the possible subsets of features

Introduction

- Hai et. al. proposed a solution to the problem of maximization of the objective functions in the CFS and mRMR measures
  - Based on polynomial mixed 0-1 fractional programming (PM01FP)

Introduction

- Here we compare CFS and mRMR solved by means of PM01FP with some feature selection measures previously used in intrusion detection
  - SVM wrapper
  - Markov blanket
  - CART (Classification and Regression Trees)

Introduction

- The comparison is practical, on a particular data set (KDD CUP ’99)
  - SVM, Markov blanket and CART were originally evaluated on that data set
- To avoid known problems with KDD CUP ’99
  - It was split into 4 parts: DoS, Probe, U2R and R2L
  - Only DoS and Probe attacks were considered, since they significantly outnumber the other 2 categories

Introduction

- Comparison by
  - The number of selected features
  - Classification accuracy of the machine learning algorithms chosen as classifiers

Feature selection methods

- Existing approaches – SVM wrapper (1)
  - A feature ranking method – one input feature is deleted from the input data set at a time
  - The resulting data set is then used for training and testing of the SVM (Support Vector Machine) classifier
  - The SVM's performance is then compared to that of the original SVM (based on all the features)

Feature selection methods

- Existing approaches – SVM wrapper (2)
  - Criteria for SVM comparison
    - Overall classification accuracy
    - Training time
    - Testing time
  - Feature ranking
    - Important
    - Secondary
    - Insignificant

Feature selection methods

- Existing approaches – Markov blanket (1)
  - Markov blanket MB(*T*) of an output variable *T*
    - A set of input variables such that all other variables are probabilistically independent of *T*
    - Knowledge of MB(*T*) is sufficient for perfect estimation of the distribution of *T* and consequently for the classification of *T*

Feature selection methods

- Existing approaches – Markov blanket (2)
  - In IDS feature selection (1)
    - A Bayesian network $B=(N,A,Q)$ from the original data set is constructed
      - N is the set of vertices – each node is a data set attribute
      - A is the set of arcs – each arc $a \in A$ represents probabilistic dependency between the attributes (variables)
      - That probabilistic dependency is quantified using a conditional probability distribution $q \in Q$ for each node $n \in N$

Feature selection methods

- Existing approaches – Markov blanket (3)
  - In IDS feature selection (2)
    - A Bayesian network can be used to compute the conditional probability of one node, given the values assigned to the other nodes
    - From the constructed Bayesian network the Markov blanket of the feature T is obtained

Feature selection methods

- Existing approaches – CART (1)
  - Classification and Regression Trees (CART)
    - Based on binary recursive partitioning
      - Binary – parent nodes are always split into exactly 2 child nodes
      - Recursive – In the next splitting, each child node is treated as a parent
    - Key elements of CART methodology
      - A set of splitting rules
      - Decision when the tree is complete
      - Assigning a class to each terminal node

Feature selection methods

- Existing approaches – CART (2)
  - In IDS feature selection
    - Contribution of the input variables to the construction of the decision tree is determined
      - By determining the role of each input variable
        - » As the main splitter
        - » As a surrogate
    - Feature importance
      - The sum across all nodes of the improvement scores

Feature selection methods

- The new approach (1)
  - A generic feature selection measure for the filter model

$$GeFS(x) = \frac{a_0 + \sum_{i=1}^{n} A_i(x)x_i}{b_0 + \sum_{i=1}^{n} B_i(x)x_i}, \quad x = (x_1, \ldots, x_n) \in \{0,1\}^n$$

  - Binary variable x_i indicates presence/absence of the feature f_i
  - A_i and B_i are linear functions of x_i

Feature selection methods

- The new approach (2)
  - The feature selection problem: find $x \in \{0,1\}^n$ that maximizes the function *GeFS*(x), i.e.

$$\max_{x \in \{0,1\}^n} GeFS(x)$$

  - Examples of instances of the *GeFS* measure
    - Correlation-feature selection (CFS)
    - Minimal-redundancy-maximal-relevance (mRMR)

Feature selection methods

- The new approach (3)
  - Correlation-feature selection (CFS)
    - Based on the average value of all feature-classification correlations and the average value of all feature-feature correlations
    - Can be expressed as an optimization problem

$$\max_{x \in \{0,1\}^n} \frac{\left(\sum_{i=1}^{n} a_{\mathrm{i}} x_i\right)^2}{\sum_{i=1}^{n} x_i + \sum_{i \neq j} 2 b_{ij} x_i x_j}$$

Feature selection methods

- The new approach (4)
  - Minimal-redundancy-maximal relevance (mRMR)
    - Relevance and redundancy of features are considered simultaneously, in terms of mutual information
    - Can be expressed as an optimization problem

$$\max_{x \in \{0,1\}^n} \left[\frac{\sum_{i=1}^{n} c_i x_i}{\sum_{i=1}^{n} x_i} - \frac{\sum_{i,j=1}^{n} a_{ij} x_i x_j}{\left(\sum_{i=1}^{n} x_i\right)^2} \right]$$

The solution

- Solving the feature selection problem (1)
  - Represent it as a polynomial mixed 0-1 fractional programming (PM01FP) task

$$\min \sum_{i=1}^{m} \frac{a_i + \sum_{j=1}^{n} a_{ij} \prod_{k \in J} x_k}{b_i + \sum_{j=1}^{n} b_{ij} \prod_{k \in J} x_k}$$

under the constraints

$$b_i + \sum_{j=1}^{n} b_{ij} \prod_{k \in J} x_k > 0, \quad i = 1, \dots, m$$

$$c_p + \sum_{j=1}^{n} c_{pj} \prod_{k \in J} x_k \le 0, \quad p = 1, \dots, m$$

The solution

- Solving the feature selection problem (2)
  - Linearize the PM01FP program to get a Mixed 0-1 Linear Programming (M01LP) problem
  - The M01LP problem can be solved e.g. by means of the branch and bound method
  - In our solution, the number of variables and constraints in the M01LP problem is linear in the number *n* of full-set features

Experimental results

- $GeFS_{CFS}$ and $GeFS_{mRMR}$ were implemented
- The goal
  - Find optimal feature subsets by means of those measures
  - Compare the obtained feature subsets with those obtained with the previously analyzed methods
    - By the cardinalities of the selected subsets
    - By accuracy of the classification

Experimental results

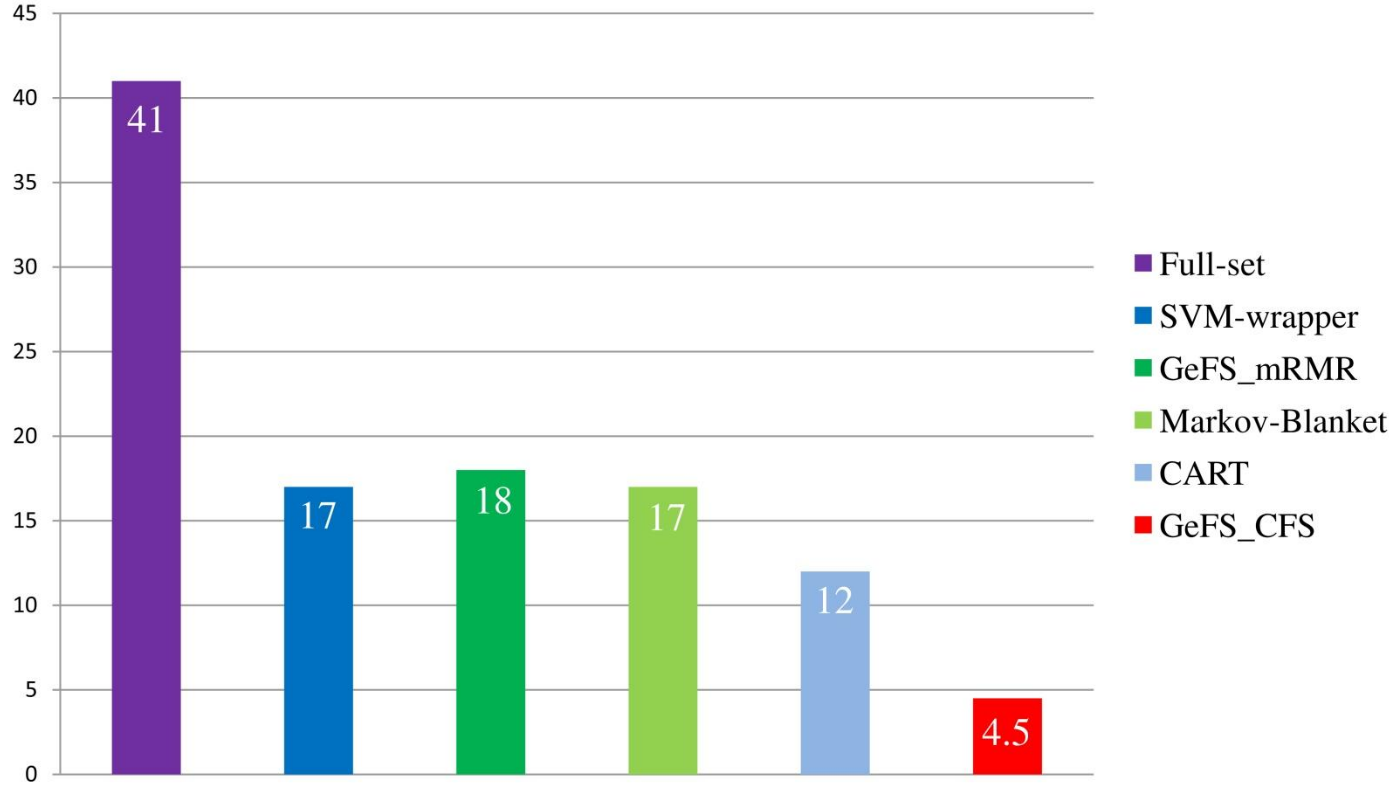
- The classification algorithm used in the experiments was the decision tree algorithm C4.5
- 10% of the KDDCUP'99 data set was used
- Only DoS and probe attacks were analyzed, for the same reason

Experimental results

- Thus, 2 data sets were generated
  - Normal traffic + DoS attacks
  - Normal traffic + probes
- Classification into 2 classes
- $GeFS_{CFS}$ and $GeFS_{mRMR}$ were run first on both data sets, to select features
- Then the classification algorithm C4.5 was run on the full-sets and the selected feature sets

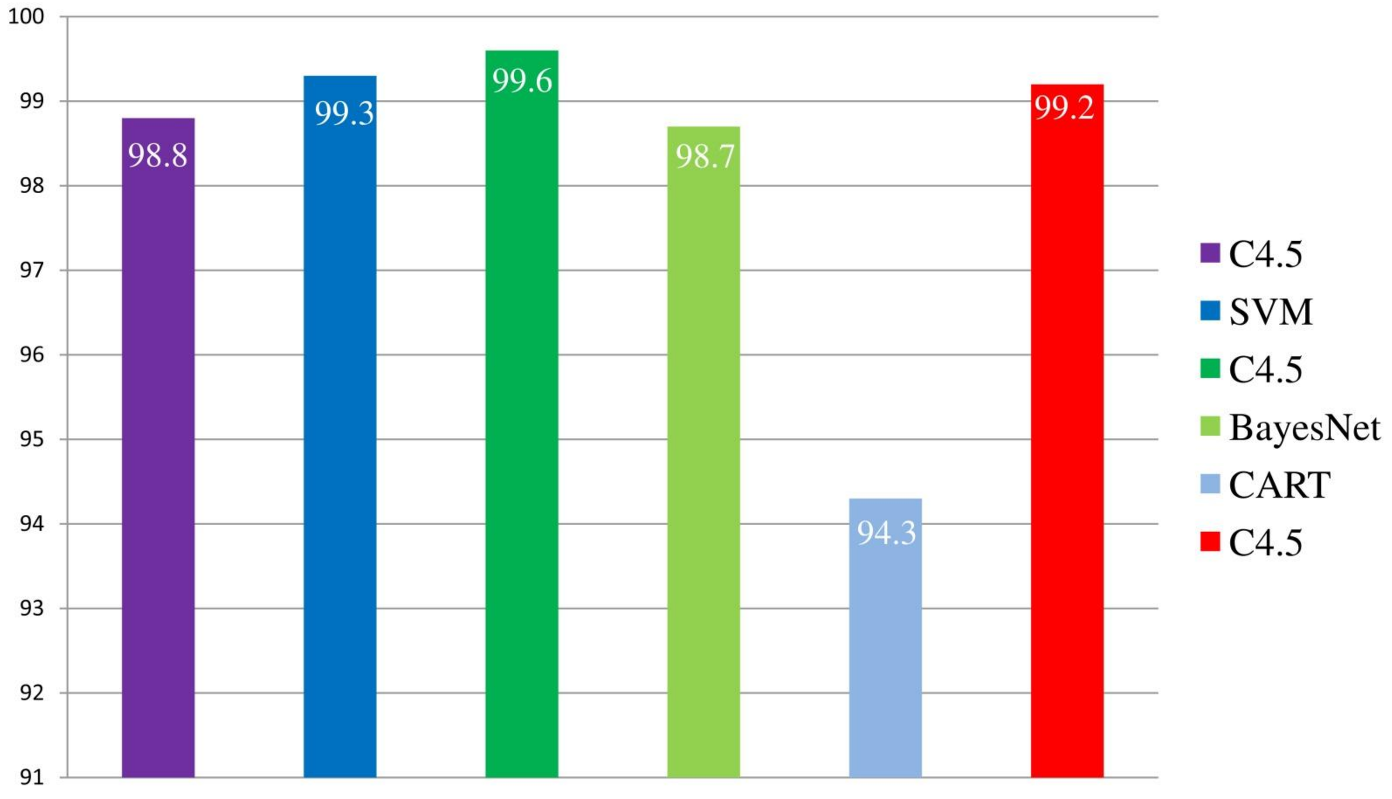
Experimental results

- The numbers of selected features (on average)



Experimental results

- Classification accuracy (on average)



Conclusions

- The *GeFS* measure instances (CFS and mRMR) performed better than the other measures involved in the comparison
  - Better (CFS) in removing redundant features
  - Classification accuracy sometimes even better and in general not worse than with the other methods