

Labelling clusters in an anomaly based IDS by means of clustering quality indexes

Roger Storløyken

Outline

- Research topic
- Research questions
- Labelling strategy
- Dunn's index
- C-index
- Experiments
- Results
- Conclusions

Research topic

- It has been shown that clustering techniques are efficient for classifying activity in an IDS
- The clustering techniques merely classify the data, without any interpretation of the nature of the data
- A major challenge is therefore to interpret the nature of obtained clusters and assign labels to them

Research topic (cont.)

- The traditional approach is cardinality based labelling strategies
- Assumption: Normal activities vastly outnumber malicious activities
- Problem: Limited capability to detect massive attacks

Research topic (cont.)

- Another labelling strategy, which solves this problem, measures the physical properties of the clusters to interpret their nature.
- Cluster quality indexes are used to control the clustering for the presence of a massive attack
- In our work we have investigated the use of Dunn's index and C-index in this strategy

Research questions

1. Can Dunn's index and C-index be applied in a labelling strategy for clustering based intrusion detection systems?
2. Which combinations of the clustering quality indexes and clustering properties yield the best accuracy of the labelling strategy?
3. What clustering quality index is best suited for labelling activity clusters, regarding accuracy and efficiency?

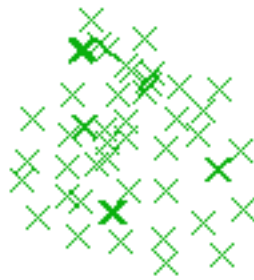
Labelling strategy

- Attack vectors corresponding to a massive attack will form very compact clusters because of their similarities, e.g. Smurf attack
- Standalone attacks are very dissimilar to each other and will form very scattered clusters
- The wide range of normal activities will form scattered clusters, but much more compact than clusters that consist of standalone attacks

Labelling strategy (cont.)



Massive attack



Normal activities



Standalone attacks

Labelling strategy (cont.)

- A hallmark of good clustering quality is compact clusters distant from each other
- This is the situation when a massive attack is present
- Clustering quality evaluation techniques (e.g. clustering quality indexes) are therefore well suited to detect massive attacks

Labelling strategy (cont.)

- A combination of evaluation techniques and e.g. the cluster diameters can therefore be used to interpret the nature of the clusters
- Example:

IF Good Clustering Quality

THEN Compact cluster = “malicious” cluster

Dunn's index

$$D = \min_{1 \leq i \leq c} \left(\min_{\substack{1 \leq j \leq c \\ j \neq i}} \left(\frac{d(c_i, c_j)}{\max_{1 \leq k \leq c} \sigma_k} \right) \right)$$

- Advantage:
 - Linear time complexity
- Disadvantages:
 - Measures only two distances
 - May be vulnerable for noise in the data

C-index

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

- Advantage:
 - Very accurate quality evaluations
- Disadvantage:
 - Quadratic time complexity

C-index (cont.)

- The C-index requires clusters of equal cardinality to produce proper quality evaluations.
- This is rarely the case in clustering based IDS
- Two modified C-indexes were developed to handle this problem; C-mean and C-small
- The modified C-indexes compute a partial C-index from both clusters in our two-cluster case

C-index (cont.)

- C-mean:
- The average of the two computations
- Advantage:
 - All distances are measured
- Disadvantages:
 - The inter- and intra-cluster distances may not be measured evenly
 - May be vulnerable for noise in the observed data

C-index (cont.)

- C-small
- Evaluation from the cluster with fewest elements
- Advantage:
 - Inter- and intra-cluster distances measured evenly
- Disadvantage:
 - Clusters with very few elements may cause unstable evaluations

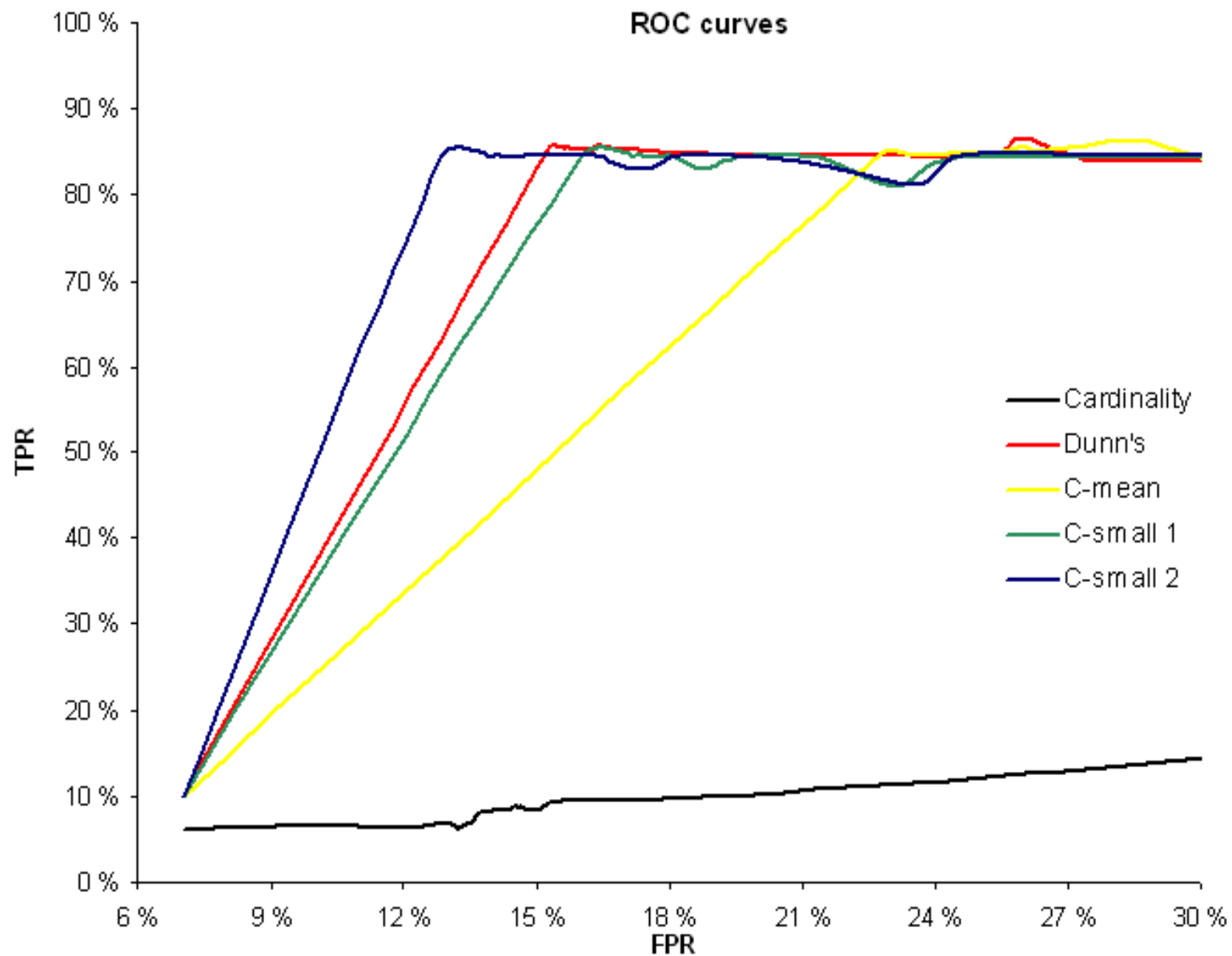
C-index (cont.)

- A very large difference between the partial indexes suggests that we have poor clustering quality
- This can be used to handle the problem with clusters that consist of very few elements when C-small is applied
- It also handles the problem with outliers (noise) in the observed data

Experiments

- Simulations with a prototype on the KDD Cup '99 data set
- Prototype:
 - Clustering based IDS with the use of K-means
 - 1000 activities are clustered into two clusters at each iteration
 - One cluster corresponds to normal traffic and one cluster to attacks
 - Dunn's index and C-index applied on top of this IDS

Results



Conclusions

- Both Dunn's index and the C-small modification can be successfully applied in the labelling strategy
- The combination of clustering quality indexes and cluster diameters yields the best labelling accuracy
- Dunn's index is best suited for situations where real-time operation is necessary.