**Data clustering for anomaly detection in NIDS[1]**

Paper này chủ yếu nói về dùng clustering trong anomaly detection. Nhược điểm lớn nhất là thời gian training và tỉ lệ false alarm cao.

Clustering dựa trên 2 assumption: normal patterns thì nhiều hơn các attack patterns và attack pattern thì không giống với các normal pattern.

Phần còn lại của paper là phần đánh giá performance bằng giải thuật k-means trên tập dữ liệu mẫu kdd Cup 1998.

(do paper dùng những thuật ngữ đã biết như “anomaly detection” “k-means” nên trong report không đề cập đến ưu nhược điểm của các giải thuật)

**Similarity Measures for Categorical Data[2]**

Mục đích đọc paper này để tìm hiểu việc clustering cho categorical data. Một số ký hiệu sử dụng trong paper:

* + Tập dữ liệu D, có N object
  + d là tập các thuộc tính
  + Ak ký hiệu cho thuộc tính thứ k
  + Giả sử rằng thuộc tính Ak gồm nk giá trị (vd: color: red, blue, green)
  + fk(x) : tần số của thuộc tính Ak có giá trị x trên tập D
  +  xác xuất thuộc tính Ak có giá trị x trên tập D
  + 

Những tính chất có thể ảnh hưởng tới các giải thuật similarity measure cho categorical data:

* + Kích thước tập D
  + Số lượng các thuộc tính
  + Số lượng giá trị của từng thuộc tính: một thuộc tính có thể có nhiều giá trị, thuộc tính khác có thể chỉ có vài giá trị. Trong paper này người ta cho rằng thuộc tính có ít giá trị hơn là quan trọng hơn các thuộc tính còn lại
  + Phân bổ tần số của các giá trị trong từng thuộc tính

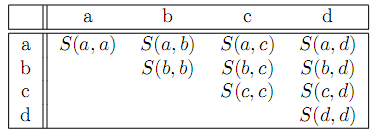
Bất kỳ giải thuật nào cũng đưa ra khoảng cách giữa 2 data instance dựa vào công thức sau:



Trong đó Sk(Xk, Yk) là similarity giữa 2 giá trị của thuộc tính Ak (chú ý rằng Xk, Yk  ∈ Ak ).

wk ký hiệu cho weight của thuộc tính Ak

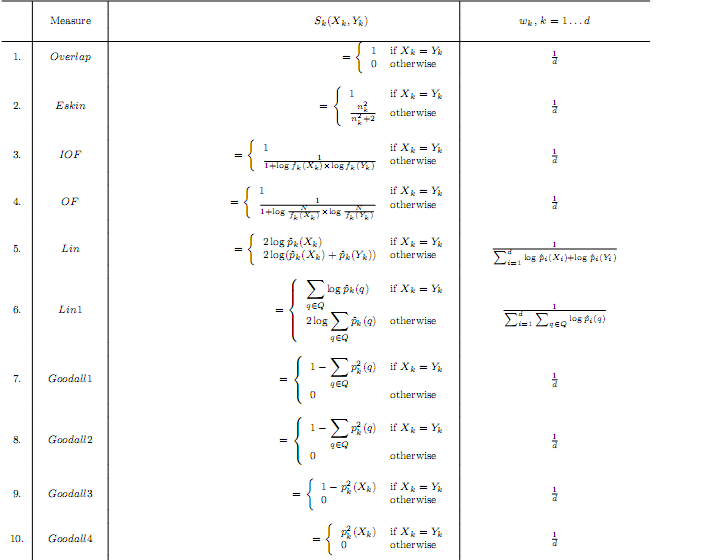
Cho 1 ví dụ sau để hiểu rõ công thức trên. Ví dụ thuộc tính A có các giá trị {a, b, c, d}

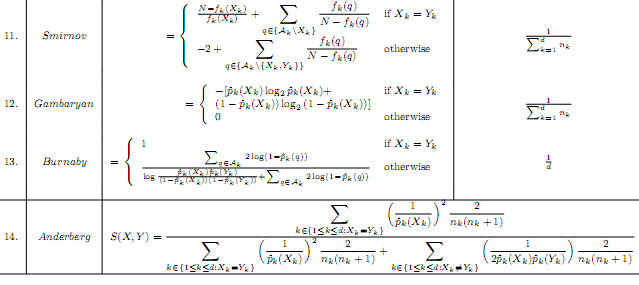


Các giải thuật similarity measure chủ yếu khác nhau ở chỗ:

* Cách tính toán các chỉ số trong matrix (như ví dụ trên S(a,a) S(a,b))
* Cách tính toán weight của từng thuộc tính

Các phương pháp similarity measure:

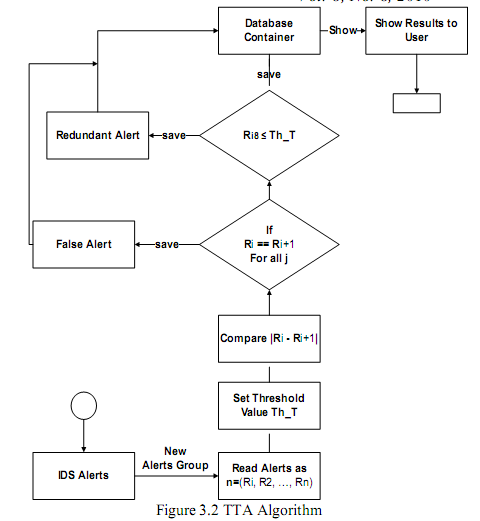




Cột thứ 3 là cách tính các chỉ số trong ma trận, còn cột thứ 4 là cách tính weight của từng thuộc tính (muốn hiểu rõ ý nghĩa của từng công thức chắc phải đọc paper của mỗi cái, không thể nói phương pháp nào tốt hơn phương pháp nào vì còn tùy thuộc vào ý nghĩa của data )

**Aggregating Intrusion Detection System Alerts Based on Row Echelon Form Concept[3**]

Paper này đề xuất một phương pháp dùng Time Threshold Aggregation (TTA) để lọc các alert trước khi đưa vào IDS. TTA dựa chủ yếu vào Row Echelon Form (chưa tìm được cái này)



Hướng lọc alert dựa vào threshold này sẽ khác hướng dùng clustering, vậy thì phương pháp nào là tốt hơn ???

Phương pháp threshold chỉ sử dụng được đối với các alert có field time, do đó không có liên quan gì lắm tới similarity measure của categorical data.

Ưu điểm của TTA:

* Hầu hết alert đều có time field
* Các ưu điểm còn lại (tương tự như ưu điểm của aggregate)

Nhược điểm: không có phần performance evaluation.

**An Anomaly Intrusion Detection System Based on Vector Quantization[4]**

Paper này trình bày ý tưởng Vector Quantization cho anomaly IDS, chủ yếu dùng để nâng cao tính hiệu quả trong việc xử lý data nhiều chiều ( tương tự như giảm chiều dữ liệu trong data mining -> không liên quan lắm tới phần của mình vì alert đã có những field nhất định không thể thu giảm được nữa)