

SISTEMAS DE DETECCIÓN DE INSTRUSOS

USANDO

TÉCNICAS DE MACHINE LEARNING

Prof. Nibaldo Rodríguez A.

OBJETIVOS

- General:
 - Implementar y evaluar un sistema de detección de intrusos (IDS) usando técnicas de machine learning.
- Específicos:
 - Seleccionar las variables explicativas relevantes usando algoritmos de Correlación Lineal, Ganancia Información, Información Mutua y Diagonalización de Matrices.
 - Calibrar los pesos de salida de una red neuronal artificial (ANN) usando el método Pseudo-inversa Moore-Penrose.
 - Calibrar los pesos ocultos de una ANN usando un algoritmo híbrido: (PSO/GSA).
 - Evaluar el rendimiento del IDS usando métricas desde la matriz de confusión.

Métodos de Selección de Variables

- Correlación Lineal (Corr)
- Ganancia de Información (IG)
- Información Mutua (IM)
- Descomposición de valores singulares (SVD)

Método #1: Corr+SVD

Calcular la Correlación de cada variable explicativa versus las clases

STEP 0:

Data: $X \in \Re^{(D \times N)}$

Clase: $Y \in \Re^{(N \times 1)}$

D : Número de Atributos (variable) de la base de datos

N : Número de muestras la base de datos

Método #1: Corr+SVD

Calcular la Correlación de cada variable explicativa versus las clases

STEP 1:

$$X \in \Re^{(D \times N)}$$

$$x, y \in \Re^{(1 \times N)}$$

$$(1) \quad x = x - \bar{x}$$

$$(2) \quad y = y - \bar{y}$$

$$(3) \quad r = \frac{\langle x, y \rangle}{\|x\|_2 \times \|y\|_2}$$

Método #1: Corr+SVD

Calcular la Correlación de cada variable explicativa versus las clases

STEP 2: Nueva Data

- Seleccionar un valor :
- (1) $\lambda \in (0,1)$
 - (2) $i = |r| > \lambda$
 - (3) $X = X(i,:)$
 - (4) $X \in \Re^{(d \times N)}$

Método#1: Corr+SVD

STEP 3: SVD

$$X \in \Re^{(d \times N)}$$

$$(1) \quad \bar{x}_i = \overline{X}(i, :), \quad i = 1, \dots, d \quad (2) \quad X = X - \bar{x}_i$$

$$(3) \quad Y = \frac{X^T}{\sqrt{N - 1}} \quad (4) \quad [U \ S \ V] = \text{SVD}(Y)$$

Método#1: Corr+SVD

STEP 3.1: Nueva Data

Seleccionar un número de variable: (1) $K \leq d$

$$(2) \quad X = V(:, 1 : K)^T \times X$$

$$(3) \quad X \in \Re^{(K \times N)}$$

Método #2: Ganancia Información +SVD

Ganancia Información : atributo versus clases

STEP 0:

Data: $X \in \Re^{(D \times N)}$

Clases: $Y \in \Re^{(N \times 1)}$

Ganancia de Información del atributo x:

$$IG(Y, x) = I(Y) - E(x)$$

D : Número de atributos (variables)

N : Número de muestras

Entropía de la Clase

STEP 1:

$$I(Y) = I(d_1, d_2, \dots, d_m) = - \sum_{i=1}^m p_i \log_2(p_i)$$

m :

Número de Clases

$$p_i = \frac{d_i}{N}$$

Probabilidad de la i -ésima Clase

d_i :

Número de muestras de la i -ésima Clases

Entropía Ponderada del Atributo:

STEP 2:

$$E(x) = \sum_{j=1}^{I_x} \frac{d_{1,j} + d_{2,j} + \dots + d_{m,j}}{N} \times I(d_{1,j}, \dots, d_{m,j})$$

I_x : Número de particiones del atributo (variable) X

$d_{i,j}$: Número de muestras de la i -ésima Clase correspondiente a la j -ésima partición

Método #2: IG +SVD

STEP 3: Nueva Data

(1) Ordenar en orden decreciente la IG obtenida previamente

$$G = \text{sort} (IG (Y, x)), \quad x \in X$$

(2) Seleccionar los top- K atributos de la base de datos

(3) Nueva Base de Datos: $X \in \Re^{(K \times N)}$

Método#2: IG+SVD

STEP 4: SVD

$$X \in \Re^{(K \times N)}$$

$$(1) \quad \bar{x}_i = \overline{X}(i, :), \quad i = 1, \dots, K \quad (2) \quad X = X - \bar{x}_i$$

$$(3) \quad Y = \frac{X^T}{\sqrt{N - 1}} \quad (4) \quad [U \ S \ V] = \text{SVD}(Y)$$

Método#2: IG+SVD

STEP 4.1: Nueva Data

Seleccionar un número de variable: (1) $d \leq K$

$$(2) \quad X = V(:, 1 : d)^T \times X$$

$$(3) \quad X \in \Re^{(d \times N)}$$

Método #3: Información Mutua (IM) +SVD

STEP 0:

Data: $X \in \Re^{(D \times N)}$

Clase: $y \in \Re^{(N \times 1)}$

D : Número de Atributos de la base de datos

N : Número de muestras la base de datos

Método #3: IM +SVD

STEP 1:

$$G(x, y) = H(x) + H(y) - H(x, y)$$

$$x, y \in R^N$$

- $H(x)$: Entropía de la variable x.
- $H(y)$: Entropía de la variable y.
- $H(x,y)$: Entropía conjunta de x e y.

Método #3: MI +SVD

STEP 2: Entropía (Shannon)

$$H(x) = -\sum_{i=1}^{I_x} p_i \log_2(p_i), \quad x \in R^N$$

$$p_i = \frac{d_i}{N}, \quad I_x = \text{ceil}\left\{\sqrt{N}; \log_2(N)\right\},$$

- I_x : Número de partición de la variable x .
- d_i : Número de muestras de la i -ésima partición de x .

Método #3: IM +SVD

STEP 3: Entropía Conjunta

$$H(x, y) = -\sum_{i=1}^{I_x} \sum_{j=1}^{I_y} p(x_i, y_j) \times \log_2(p(x_i, y_j)) \quad x, y \in R^N$$

$$p(x_i, y_j) = \frac{d_i \times d_j}{N} \quad I_x, I_y = \{\sqrt{N}; \log_2(N)\}$$

- I_x, I_y : Número de particiones de las variables x e y.
- d_i : Número de muestras de la i -ésima partición de x .
- d_j : Número de muestras de la j -ésima partición de y .

Método #3: IM +SVD

STEP 4: Correlación

$$R(x, y) = 2 \times \frac{G(x, y)}{H(x) + H(y)}$$

Método #3: IM +SVD

STEP 5: Nueva Data

(1) Ordenar en orden decreciente los valores R obtenidos previamente

$$R_s = \text{sort} (R(Y, x)), \quad x \in X$$

(2) Seleccionar los top-K atributos de la base de datos

(3) Nueva Base de Datos: $X \in \Re^{(K \times N)}$

Método #3: IM +SVD

STEP 6: SVD

$$X \in \Re^{(K \times N)}$$

$$(1) \quad \bar{x}_i = \overline{X}(i, :), \quad i = 1, \dots, K \quad (2) \quad X = X - \bar{x}_i$$

$$(3) \quad Y = \frac{X^T}{\sqrt{N - 1}} \quad (4) \quad [U \ S \ V] = \text{SVD}(Y)$$

Método#2: IG+SVD

STEP 6.1: Nueva Data

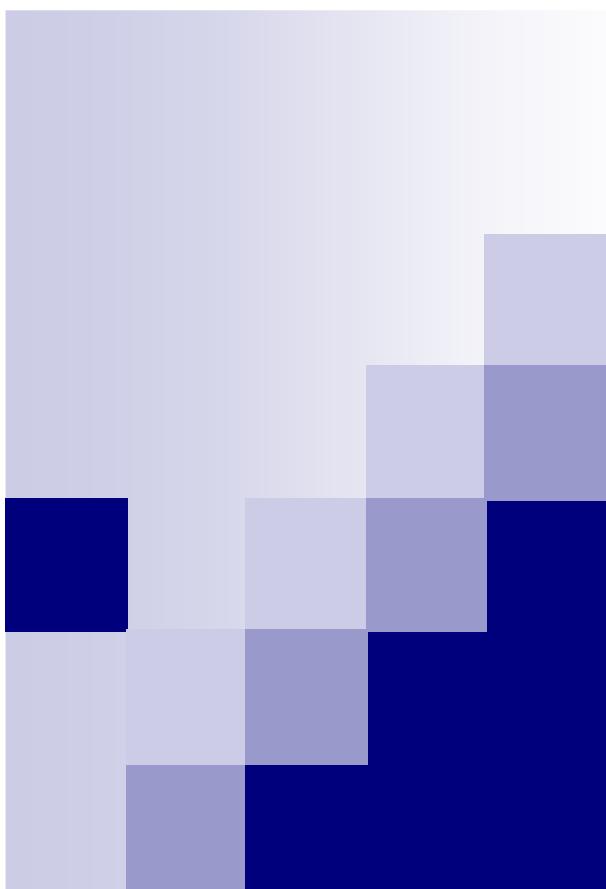
Seleccionar un número de variable: (1) $d \leq K$

$$(2) \quad X = V(:, 1 : d)^T \times X$$

$$(3) \quad X \in \Re^{(d \times N)}$$

MACHINE LEARNING (ML)

Prof. Nibaldo Rodríguez A.



CONTINUARÁ....