**BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA**

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

**Clasificador Basado en Centroides**

**EQUIPO:**

*Alvarez Santos Fredy*

*Flores Velázquez David*

*Juárez Escárcega Ilse Lilian*

*Torres Timal Martín Rafael*



*Procesamiento del Lenguaje Natural*

*Dr. David Eduardo Pinto Avendaño*

***Otoño 2012***

Clasificador basado en Centroides

Existen multitud de clasificadores diferentes. Algunos se basan en el concepto de distancia entre los vectores de características. Sin embargo, hay que decir que existen muchos otros enfoques. A continuación, se expone un clasificador basado en distancia, que permiten por su simplicidad y variedad una aproximación didáctica y adecuada al problema de la clasificación.

**Algoritmo**

El algoritmo consiste en lo siguiente, dado un corpus de entrenamiento con la siguiente estructura:

*Idi, ClaseX, atributo1, atributo2, … , atributoN*

*…*

*IdM, ClaseX, atributo1, atributo2, … , atributoN*

Se procede a que para cada objeto de la ClaseX en el corpus, calcular la media de los atributos:

*MediaObjetoX = (atributo1, +atributo2 + … + atributoN)/N*

Posteriormente calcular la media General de las medias de los objetos de la Clase

*MediaClaseX = (MediaObjeto1X + MediaObjeto2X + … + MediaObjetoMX)/M*

Donde M es el numero de elementos encontrados pertenecientes a una cierta clase.

**Código**

awk '{

$2 = tolower($2);

sum[$2]=0;

for (i=3; i<=NF; i++){

sum[$2] = sum[$2] + $i; #sumar los N atributos por elemento

}

sum[$2]=sum[$2]/(NF-2); #tomar el promedio parcial de los atributos

media[$2]=media[$2]+sum[$2]; #sumatoria de los promedios parciales

Total[$2]++; #total por de elementos por Clase

}

END {

for (x in media)

print x, media[x]/Total[x]; #Para cada Sumatoria de los promedios

} ' $\* #obtenemos la media General de cada Clase

Esto Generara como salida un modelo el cual tendrá la siguiente estructura:

*ClaseX , Centroide*

*…*

*ClaseN , Centroide*

Donde Centroide = MediaClaseX

Ejemplo Modelo

*lingustrumvulgare 5.425*

*ficusbenjamina 5.05*

*popolusalba 5.525*

*…*

Una vez Generado el Modelo ya podemos Clasificar a que clase pertenece una nueva entrada de la siguiente manera:

**Algoritmo**

*DistanciaMinima=infinito;*

*ClaseResultado;*

*Para cada X in Clases{*

*sum=0;*

*for (i=1; i<Numero de Atributos; i++) {*

*sum=sum + Atributoi;*

*}*

*sum = sum/(Numero de Atributos); #hasta este punto calculamos el*

*#promedio de los atributos del nuevo*

*# elemento a clasificar.*

*distancia = abs(CentroideX - sum) #se calcula la distancia al centroide X*

*if ( distancia < DistanciaMinima) { #Checamos si la distancia es menor a*

*# la DistanciaMinima actual*

*DistanciaMinima= distancia; #guardamos la nueva distancia*

*ClaseResultado= x; #guardamos la clase resultado*

*}*

*}*

**Código**

awk '

FILENAME==model{

modelo[$1] = $2;

clases[$1] = 2;

next;

}

function abs(x){

return (((x < 0.0) ? -x : x) + 0.0)

}

BEGIN {

EPSILON = 0.00001;

}

{

minimo = 999999999999999999999999999;

for (x in clases) {

sum=0;

for (i=2; i<NF; i++) {

sum=sum + $i;

}

sum = sum/(NF-1);

dif = abs(modelo[x]-sum);

if ( dif < minimo) {

minimo = dif;

clase = x;

}

}

print "--------------------------------------------------------";

print $1, clase, minimo, " : ", $2, $3, $4, $5"...";

print "--------------------------------------------------------";

}

END{

}' model=$1 $\*