I. Definiciones

Definiciones formales usadas en los macro-algoritmos presentados en la parte II

*** Random Forest**

(1) Gini = 1 - $\sum_{i=1}^{C} (p_i)^2$

Donde:
C = Clases
p = probabilidad de la clase

*** LDA**

(2) μ = $\frac{1}{n}\sum_{i=1}^{n} x_i$

(3) C = $\frac{1}{n_1+n_2}(n_1 C_1 + n_2 C_2)$

(4) β = $C^{-1}(\mu_1 - \mu_2)$

Donde:
μ = media de las características
C = Co-varianza
n = número de características
x_i = característica i
β = importancia de característica

*** F-meassure**

precision = $\frac{True\ Positive}{True\ Positive + False\ Positive}$

recall = $\frac{True\ Positive}{True\ Positive + False\ Negative}$

(5) F_1 = $2 * \frac{precision * recall}{precision + recall}$

*** Accuracy**

(6) accuracy_score = $\frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$

*** Mayoría**

(7) si C_i >= num_feature_sets / 2 entonces:
contador = contador + 1

Donde:

C_i = Característica i
num_feature_sets = Cantidad de feature_sets total generadas en el proceso de preselección.

*** Calidad-Cantidad**
(8) P = α / num_features_total

Donde:
α = valor máximo de penalización (en caso de usar todas las características del dataset)

(9) score_final = score - (P * len(feature_set))

Donde:
score = resultado de la clasificación con radom forest o LAMDA-HAD
P = penalización calculada en (8)
len(feature_set) = longitud del feature_set actual.

II. Macro Algoritmo

El macro-algoritmo consta de 3 partes:

Proceso de preselección:

En esta fase se extraen los primeros conjuntos de características usando Random Forest y LDA

Entrada: Conjunto de datos (dataset)
Salida: Diferentes conjuntos de características seleccionadas (feature set)

* Definir el modelo de Random Forest (RF) a usar.
Entradas:
- Número de árboles (n_estimators)
- Nivel máximo de profundidad de cada árbol (max_depth)
- Número de características a considerar al buscar la mejor ramificación en cada árbol (max_features)
- Valor de umbral (threshold)

(Estos son valores definidos por el usuario)

* Entrenar el modelo con el dataset.
1. Desde 0 hasta n_estimators:
1.1 Elegir características aleatorias en el dataset en un rango de max_features
1.2 Construir un árbol de decisión con esas características y una profundidad de max_depth usando (1)
1.3 Calcular la importancia de cada característica promediando los valores obtenidos en el paso 1.2
2. Calcular las importancias totales promediando las importancias calculadas en cada árbol del paso 1
3. Seleccionar características cuyas importancias sean mayores al valor de umbral

* Definir el modelo de LDA a usar:

Entradas:

- Valor de umbral (threshold)

* Entrenar el modelo con el dataset.

1. Calcular medias de las características usando (2)
2. Calcular co-varianzas de las características usando (3)
3. Calcular importancias de las características usando (4)
4. Seleccionar características cuyas importancias superen valor de umbral

* Cada modelo genera diferentes conjuntos de características en base a diferentes valores de parámetros usados

Proceso de combinación:

En esta fase se realiza una combinación de los conjuntos de características extraídos en la fase anterior usando diferentes criterios de combinación y selección.

Entrada: Diferentes conjuntos de características generados en proceso de preselección
Salida: Conjuntos de características combinados o seleccionados

1) Selección en base a calidad de clasificación:

Este criterio utiliza los resultados de precisión en el proceso de clasificación de los clasificadores Random Forest y LAMDA-HAD para seleccionar el sub-conjunto de características con los resultados mas altos.

Macro-algoritmo:

* Por cada conjunto de características obtenido del proceso de preselección se realiza un proceso de clasificación con los clasificadores LAMDA-HAD y RF.

* Calcular la precisión de los resultados con f-measure usando (5) y accuracy usando (6).

* Seleccionar el conjunto de características que haya obtenido el mejor resultado.

2) Combinación completa:

En este caso se realiza una combinación completa de todas las características extraídas en la primera fase en un único conjunto de características.

Macro-algoritmo:

* Iterar por cada conjunto de características de entrada.

* Si la característica de la iteración actual no se encuentra en el conjunto de características de salida agregarla.

* Finalizar con un conjunto de características que combina todas las características generadas en el proceso de preselección.

3) Votación por mayoría:

Este criterio examina los conjuntos de características extraídos en la primera fase y selecciona aquellos que se encuentren en mas de la mitad de esos conjuntos generados.

Macro-algoritmo:

* Iterar por cada conjunto de características de entrada.

* Establecer un contador por cada característica encontrada.

* Incrementar el contador usando la formula (7)

* Finalizar con un conjunto de características que agrupa aquellas con mayor aparición en todos los conjuntos de entrada.

4) Selección calidad-cantidad:

Este criterio también utiliza la precisión obtenida al aplicar los clasificadores Random Forest y LAMDA-HAD pero además asigna valores de penalización a cada conjunto por cantidad de características que contenga, es decir, a mayor número de características en el conjunto, mayor penalización y menor oportunidad de ser seleccionado.

Macro-algoritmo:

* Por cada conjunto de características obtenido del proceso de preselección se realiza un proceso de clasificación con los clasificadores LAMDA-HAD y RF.

* Calcular la precisión de los resultados con f-measure usando (5) y accuracy usando (6).

* Calcular valor de penalización por característica usando (8)

* Aplicar penalización al resultado obtenido en la clasificación usando (9)

Proceso de postselección:

Esta fase final realiza un último cálculo de la precisión usando los clasificadores obtenidos en la fase anterior y seleccionando el conjunto con mayor valor como el conjunto de características óptimo final.

Entrada: Conjuntos de características generados en el proceso de combinación
Salida: Conjunto de características final

* Por cada conjunto de características obtenido del proceso de combinación se realiza un proceso de clasificación con los clasificadores LAMDA-HAD y RF.

* Calcular la precisión de los resultados con f-measure usando (5) y accuracy usando (6).

* Se selecciona el conjunto de características con valores de precisión mas altos.