Ejercicios de la clase 5: "Support Vector Machines"

El conjunto de datos contenido en el archivo `Refimprove_unbalanced.arff` en formato WEKA, posee 1000 artículos de Wikipedia en español que presentan la falla *Referencias Adicionales*[1] y representan nuestra clase positiva. Asimismo, como clase negativa se tienen 12000 artículos de Wikipedia que no han sido etiquetados; es decir, que pueden contener la falla o no. Como puede observarse, este problema de clasificación es desbalanceado, pues no se tiene la misma cantidad de muestras de cada clase que se quiere aprender a distinguir.

Se pide:

1. Ejecutar el código `curso_MD_cross_val_SVM_grid_search.java` para realizar una búsqueda en grilla para un kernel lineal, con $C \in \{2^5, 2^7, 2^9, 2^{11}\}$, para el dataset original contenido en `Refimprove_unbalanced.arff`.

   a) Asumiendo que ya se ha modificado el código `.java` convenientemente y se lo ha compilado, ubicados en el directorio `/mineria-de-datos/miscelaneos/`, un ejemplo de ejecución sería:

   ```
   java -cp ./libsvm.jar:./weka.jar:. curso_MD_cross_val_SVM_grid_search
   ../colecciones_weka/Refimprove_unbalanced.arff > lineal.unb.out
   ```

   b) Dada la cantidad de artículos contenidos en el dataset la ejecución puede demorar. En una prueba realizada, la ejecución descrita en el punto anterior, demoró 156 minutos en una notebook con un procesador Intel Core i5 de 4 núcleos y 4Gb de RAM.

   c) En `lineal.unb.out`, registrar el valor de C que ha obtenido la mejor medida F (`F-Measure`) para la clase positiva (`Refimprove`).

2. Balancear el dataset original contenido en `Refimprove_unbalanced.arff` de manera que ahora sólo hayan 1000 artículos de la clase `untagged`. Se sugiere que estos 1000 artículos sean seleccionados siguiendo una distribución uniforme. No obstante esto, describa la forma en que finalmente se procedió para balancear el dataset.

   a) Ejecutar nuevamente el código `curso_MD_cross_val_SVM_grid_search.java` para realizar una búsqueda en grilla para el dataset balanceado con kernel lineal con $C \in \{2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}, 2^{15}\}$.

   b) Dado que la cantidad de artículos contenidos en el dataset ha disminuido, si bien se han agregado nuevos valores de C a evaluar, el tiempo de ejecución decrece notablemente. En una prueba realizada, la ejecución descrita en el punto anterior, demoró 6 minutos en una notebook con un procesador Intel Core i5 de 4 núcleos y 4Gb de RAM.

   c) Registrar el valor de C que ha obtenido la mejor medida F (`F-Measure`) para la clase positiva (`Refimprove`).

   d) Responda:

      1) ¿Con qué conjunto de datos (balanceado vs. desbalanceado) se obtuvo un mejor valor de medida F? ¿Era esperable ese resultado en función de cómo se procedió a balancear el dataset?

[1] `https://es.wikipedia.org/wiki/Plantilla:Referencias_adicionales`

2) Complemente la justificación de su respuesta del punto anterior, incluyendo una breve mención sobre los valores de C hallados para cada enfoque.
3) Para los casos en que los valores de C coincidieron, es decir: $C \in \{2^5, 2^7, 2^9, 2^{11}\}$, ¿qué conjunto de datos permitió entrenar un mejor clasificador lineal?

3. Ejecutar nuevamente el código `curso_MD_cross_val_SVM_grid_search.java` para realizar una búsqueda en grilla para el dataset balanceado diseñado en el punto 2 con un kernel RBF con $C \in \{2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}, 2^{15}\}$ y $\gamma \in \{2^{-1}, 2^1, 2^3\}$.

   a) Registrar la mejor combinación de valores de C y γ que hayan obtenido la mejor medida F (`F-Measure`) para la clase positiva (`Refimprove`).

   b) ¿Tuvo el kernel RBF una mejor performance que los clasificadores lineales entrenados anteriormete? Justifique su respuesta.

4. Desde el explorador gráfico de WEKA, evalúe manualmente un kernel lineal, con $C \in \{2^5, 2^7, 2^9, 2^{11}\}$, para el dataset original cotenido en `Refimprove_unbalanced.arff`, pero esta vez, penalizando los artículos mal clasificados de ambas clases de forma diferente. Para ello, como se muestra en la Figura 1 se deben introducir separados por un espacio en blanco, el peso para penalizar la clase postiva, y luego el de la clase negativa. Como se muestra en la figura se sugiere $w_+ = 6$ y $w_+ = 1$. En la Figura 2, se puede apreciar cómo ha aparecido la opción `-W "6.0 1.0"` en el string de formato.

5. Para cada ejecución del punto anterior, registrar la mejor combinación de valor de C con los pesos sugeridos, que hayan obtenido la mejor medida F (`F-Measure`) para la clase positiva (`Refimprove`). ¿Qué puede concluir sobre cómo ponderar las clases ha impactado en el valor de C al evaluar el clasificador usando la medida F (`F-Measure`) para la clase positiva (`Refimprove`)?

6. Para el mejor clasificador obtenido en cada uno de los puntos 1.a, 2.a, 3.a y 4, utilice el conjunto de test provisto en el archivo `Refimprove.test.csv.arff` para evaluar la medida F (`F-Measure`) para la clase positiva (`Refimprove`). Para el caso de los clasificadores de los puntos 1.a, 2.a, 3.a, hágalo ejecutando el código `curso_MD_test_set_SVM.java`, modificando convenientemente los parámetros de cada clasificador. Para el clasificador 4, siga trabajando en modo gráfico y seleccione el conjunto de test como muestra la Figura 3.

7. ¿Los resultados obtenidos por cada clasificador sobre el conjunto de test, fue análogo a la performance obtenida en la etapa de cross-validation cuando se determinó cual parecía ser la mejor configuración? Comente al respecto y justifique sus resultados.

8. El conjunto de test `Refimprove.test.csv.arff` está balanceado, mientras que los clasificadores entrenados en los puntos 1.a y 4, fueron entrenados con un conjunto de datos desbalanceados. Conjeture qué sucedería en cuanto a la performance que exhibiría cada clasificador si el conjunto de test tuviera el mismo grado de desbalance en las clases.

9. Construya a partir del conjunto `Refimprove.test.csv.arff`, un conjunto de test desbalanceado que respete la proporción existente entre artículos `Refimprove` y `untagged` en `Refimprove_unbalanced.arff`. Finalmente, evalué el mejor clasificador obtenidos en el punto 1.a y en el punto 4, y corroboré sus conjeturas. ¿Fueron acertadas? Justifique.

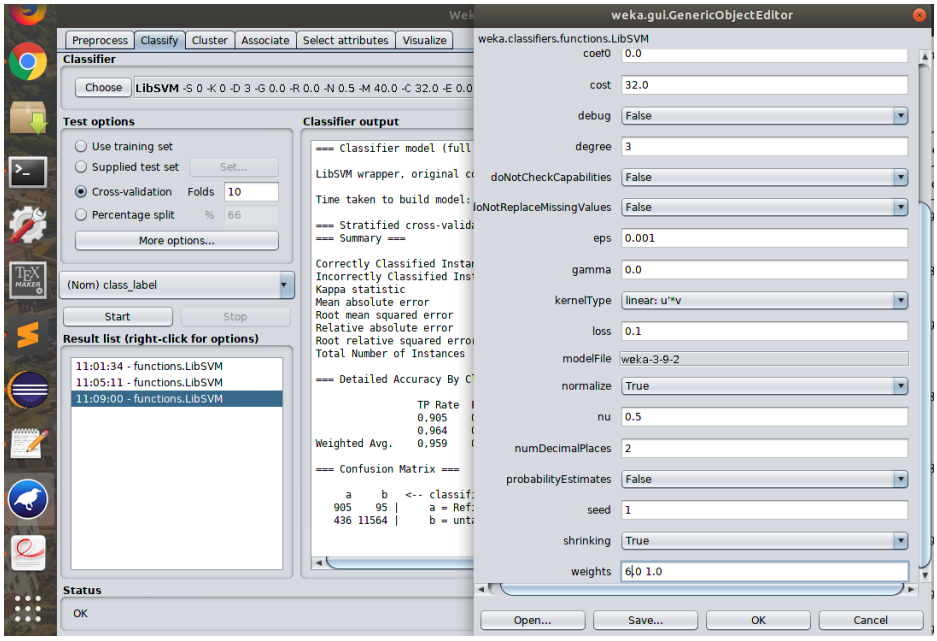

Figura 1: Completando los pesos para penalizar cada clase de forma diferencial

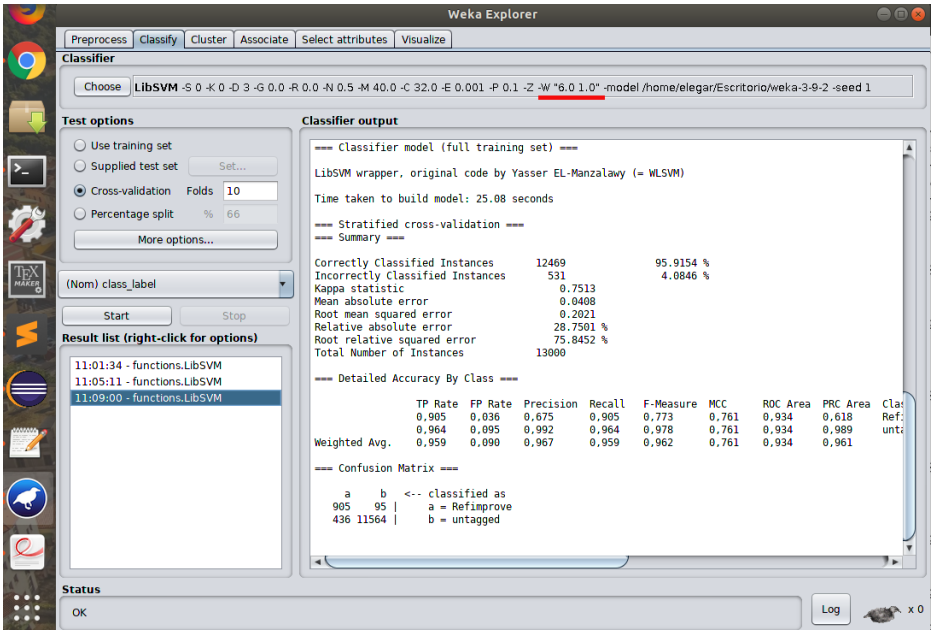

Figura 2: Pesos ya incorporados en el string de formato de parámetros que se le pasa al clasificador

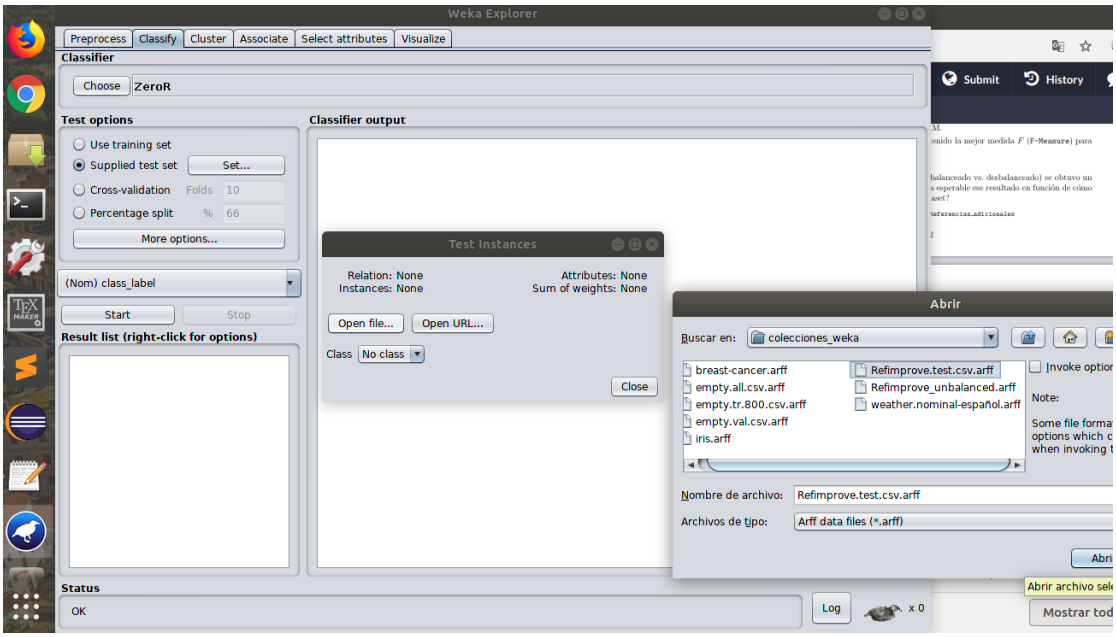

Figura 3: Conjunto de test para evaluar de forma equitativa a todos los enfoques