CLASIFICACIÓN DE DOCUMENTOS Y CURVAS ROC

Práctica 3

Los conjuntos de datos ReutersCorn-train.arff y ReutersGrain-train.arff son conjuntos de datos de entrenamiento derivados de colecciones de artículos que se utilizan como referencia para evaluar clasificadores de documentos. ReutersCorn-test.arff y ReutersGrain-test.arff son sus correspondientes conjuntos de prueba.

Los documentos en los conjuntos Corn y Grain son los mismos. Solo difieren en las etiquetas de clase. En el primer conjunto de datos, los artículos relacionados con Corn tienen el valor de clase 1 y los restantes 0. El objetivo es construir un clasificador que identifique artículos relacionados con Corn. En el segundo conjunto de datos las etiquetas se elaboran para los artículos relacionados con Grain.

Ejercicio 1: Crear clasificadores para los dos conjuntos de datos, con *FilteredClassifier*, aplicando *StringToWordVector* con J.48 y NBMultinomial, evaluándolos sobre el correspondiente conjunto de test.

1-a.¿Qué porcentaje de clasificación correcta se obtiene en los cuatro escenarios?

Tasa de acierto %		
	J48	**NBMultinomial**
Corn	97.351	93.709
Grain	96.358	90.728

1-b. En base a las tasas de error, ¿qué clasificador elegiría para cada conjunto de datos?

Corn

Para este conjunto de datos, con el clasificador *J48* obtenemos una tasa de error del 2.649% mientras que con el clasificador *NBMultinomial* obtenemos una tasa de error del 6.291 %. Como se puede observar, la tasa de error es ligeramente inferior para el clasificador **J48**.

Grain

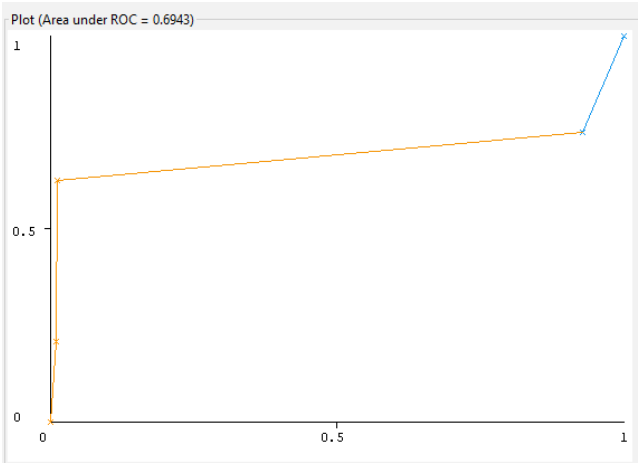
Para este conjunto de datos, con el clasificador *J48* obtenemos una tasa de error del 3.642% mientras que con el clasificador *NBMultinomial* obtenemos una tasa de error del 9.272 %. En este caso, la diferencia es un poco mayor que en el caso anterior, sin embargo, la tasa de error es ligeramente inferior para el clasificador **J48**.

Ejercicio 2: En la tabla *Detailed Accuracy by Class* se calcula el área bajo la curva ROC, AUC. Weka elabora la curva ROC de cada clase, que denomina Threshold curve. Se visualiza pinchando sobre el último experimento realizado con el botón derecho.

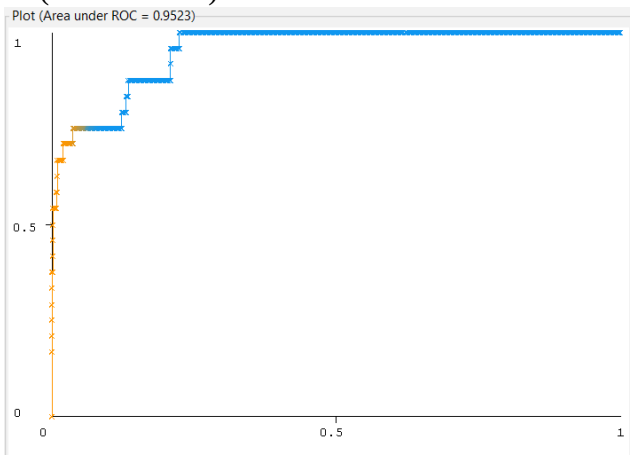
2-a.Comparar las curvas ROC de ambos clasificadores para la clase de interés en cada conjunto de datos.

Corn

- J48 (AUC = 0.694)



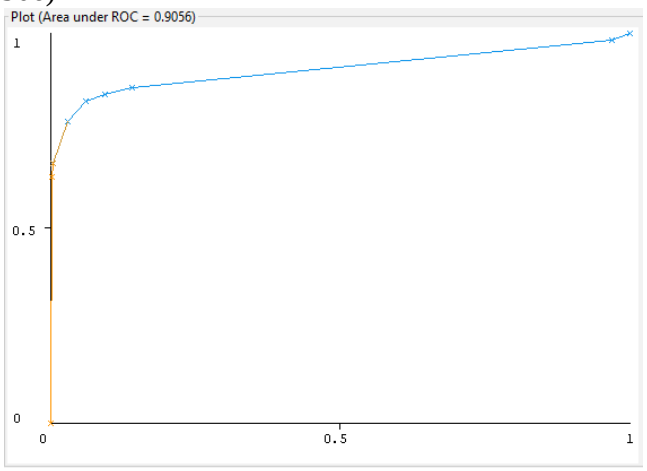
- NBMultinomial (AUC = 0.952)



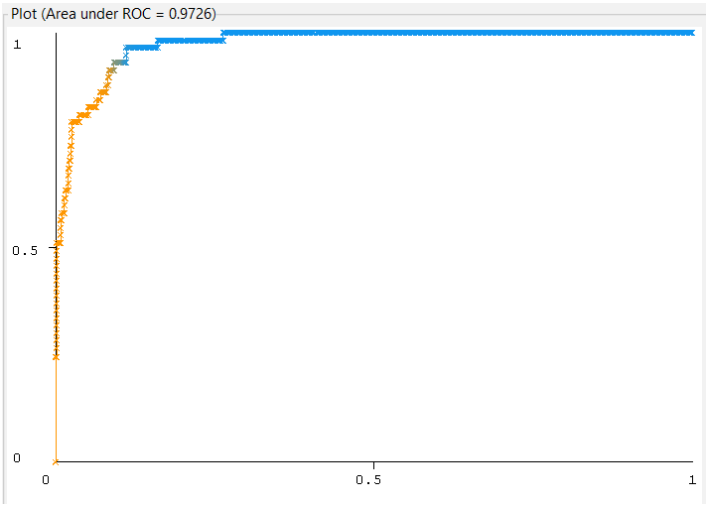
Como se puede observar, el área bajo la curva ROC que genera el clasificador NBMultinomial (0.952) es mayor que la generada por J48 (0.694).

Grain

- J48 (AUC = 0.906)



- NBMultinomial (AUC = 0.973)



Como se puede observar, el área bajo la curva ROC que genera el clasificador NBMultinomial (0.973) es mayor que la generada por J48 (0.906).

2-b. En base a las curvas ROC, ¿qué clasificador elegiría para cada conjunto de datos?

Como se ha comentado, para el conjunto de datos **Corn** el clasificador **NBMultinomial** generaba un AUC mayor, por lo que elegimos este clasificador.

Para el conjunto de datos **Grain** el clasificador **NBMultinomial** vuelve a generar un AUC ligeramente superior, por lo que elegimos este clasificador.

Ejercicio 3: En la clasificación de documentos se utilizan otras métricas para evaluar los clasificadores, como *precisión* y *recall*, o la *medida-F*. Todas ellas se elaboran a partir de TP, FP, TN, FN. Sus valores se incluyen en la tabla *Detailed Accuracy by Class*.

En base a sus definiciones, ¿cuáles son los mejores posibles valores para estas métricas? Describir en qué circunstancias se obtiene estos mejores valores.

Precision: porcentaje de instancias recuperadas que son relevantes *(TP / (TP+FP))*. Si este valor es próximo a 1, entonces se entiende que todos los documentos que se recuperan son relevantes. Cuantos menos *falsos positivos* (FP) se detecten, mayor será la precisión.

Recall: porcentaje de distancias relevantes que han sido recuperadas *(tp = TP / P)*. Si este valor es próximo a 1, entonces se entiende que se recuperan todos los documentos que son relevantes. El mejor valor para esta métrica se da cuando el número de ciertos positivos sea próximo al total de positivos clasificados.

Measure: medida de precisión de un test que se obtiene ponderando la precisión y el recall, en el caso de esta fórmula, estamos dando la misma importancia a ambos valores *((2*Recall*Precision) / (Recall+Precision)).* Cuanto mayores sean los valores de precisión y recall, más preciso será el test que se realiza.

Corn			
	Precision	***Recall***	***F-Measure***
J48	0.682	0.625	0.652
NBMultinomial	0.360	0.750	0.486

Grain			
	Precision	***Recall***	***F-Measure***
J48	0.927	0.667	0.776
NBMultinomial	0.505	0.912	0.650

Para el conjunto de datos Corn, se obtienen mejores métricas con el clasificador J48 tanto para *Precision* como para *F-Measure,* aunque para *Recall* se obtiene mejor valor con el clasificador NBMultinomial.

Por último, para el conjunto de Grain ocurre exactamente lo mismo.