Master Universitario en Bioinformática para Ciencias de la Salud
Facultad de Informática
UNIVERSIDADE DA CORUÑA

Informe de la Práctica 5.

INTELIGENCIA COMPUTACIONAL PARA BIOINFORMÁTICA.

Pedro Sánchez García
MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA PARA CIENCIAS DE LA SALUD.
PROFESORA: DRA. NOELIA SÁNCHEZ MAROÑO.

PRÁCTICA 5. REDUCCIÓN DE LA DIMENSIÓN

CONJUNTO DE DATOS: QSAR BIODEGRADATION

PREPROCESADO DE DATOS: NORMALIZACIÓN.

En la presente práctica, nos centramos en la reducción de la dimensión. Para llevar a cabo los análisis con diferentes metodologías, se toma como punto de partida el conjunto de QSAR biodegradation normalizado en las anteriores prácticas, con el fin de evitar posible dispersión entre los valores de las variables y de sesgos en el modelo de aprendizaje empleado.

APRENDIZAJE.

CONFIGURACIÓN DEL MÉTODO DE APRENDIZAJE.

Tal y como hemos visto en las anteriores prácticas, el conjunto de QSAR conforma un problema de clasificación binaria con 41 variables. Ante este amplio número de características, se puede producir un sobreajuste en el método de aprendizaje. En consecuencia, es preciso eliminar información que aporta redundancia en el conjunto de datos, basándonos en unas dimensiones menores del espacio de entrada. De esta forma, se logra una menor complejidad en el problema tratado y, por tanto, un menor tiempo de ejecución necesario.

Como aproximaciones, debemos tener en cuenta que se distinguen: (1) Selección de características: Se determina un subconjunto de variables originales (Ej.: Método Relief). (2) Extracción de características: Se basa en la combinación de variables originales para generar nuevas variables, con el objetivo de perder menor información y obtener problemas más sencillos (Ej.: Análisis de Componentes Principales).

Para esta práctica, se emplea la técnica del análisis de componentes principales como extracción de características, cuyo fundamento consiste en la proyección de los datos en un nuevo subespacio, manteniendo la información relevante en mayor medida, es decir, maximizando la varianza. Por tanto, se elige para esta última cuestión un valor mínimo de 90 en la configuración. En lo que respecta a la versión de red neuronal, se usa la función de transferencia lineal, con 84 unidades en la capa oculta, entrenando con las entradas transformadas y *targets* correspondientes.

La aproximación para seleccionar características se lleva a cabo por el método Relief, que tiene como base la ausencia de heurística y tolerancia al ruido. En general, el algoritmo asigna pesos a las características, tomando una muestra al azar y buscando otra de la clase contraria. Con esto, se evalúa la diferencia existente en clases para las características y se puede apreciar si estas últimas contribuyen a la distinción entre clases. La versión de red neuronal se mantiene con los mismos parámetros empleados en la otra aproximación.

MÉTODO DE ESTIMACIÓN DEL ERROR: VALIDACIÓN CRUZADA.

Tal y como se ha establecido en las anteriores prácticas, dado el tamaño muestral del conjunto de datos de QSAR, que es superior a 1000 muestras, se elige la validación cruzada como el método para estimar el error. Se divide el conjunto de datos en k=5 subconjuntos disjuntos con un tamaño similar. Progresivamente, se seleccionan 4 subconjuntos para el entrenamiento o proceso de aprendizaje y uno de los subconjuntos para el test, mediante el cual se estima el error real del modelo. De este modo, se trata de un proceso con k = 5 iteraciones y 3 repeticiones.

MEDIDAS DE RENDIMIENTO.

Para las diferentes aproximaciones tratadas, se han calculado las siguientes medidas de rendimiento: sensibilidad (recall), especificidad, precisión, valor predictivo negativo, exactitud (accuracy) y F1-Score. A continuación, se muestran los resultados de cada medida para entrenamiento y test en cada aproximación, lo que nos permitirá su comparación con las técnicas de las prácticas anteriores:

EXTRACCIÓN DE CARACTERÍSTICAS: *ANÁLISIS DE COMPONENTES PRINCIPALES*

Tabla 1. Medidas de rendimiento global para entrenamiento y test con la técnica de análisis de componentes principales.

	Entrenamiento	Test
Sensibilidad (Recall)	0,8267 ± 0,1611	0,8157 ± 0,1533
Especificidad	0,8267 ± 0,1611	0,8157 ± 0,1533
Precisión	0,8499 ± 0,0633	0,8314 ± 0,0675
Valor predictivo negativo	0,8499 ± 0,0633	0,8314 ± 0,0675
Exactitud (accuracy)	0,8533 ± 0,0489	0,8414 ± 0,0473
F1-Score	0,8223 ± 0,1499	0,8116 ± 0,1371

Los resultados de las medidas de rendimiento son similares para entrenamiento y test de forma global (Tabla 1). No obstante, se puede observar que las medidas en el caso del entrenamiento no presentan valores próximos a 1, lo que pone de manifiesto que no se produce una generalización adecuada y, en consecuencia, un entrenamiento adecuado de la red. En test, esto conduce a unos valores de las medidas que también se distancian de 1.

Si nos centramos en la sensibilidad y especificidad, no podemos afirmar que la versión de red, con las entradas transformadas, proporcione un buen comportamiento para la clasificación de los compuestos químicos.

SELECCIÓN DE CARACTERÍSTICAS: *RELIEF*

Tabla 2. Medidas de rendimiento global para entrenamiento y test con la selección de características mediante el algoritmo Relief.

	Entrenamiento	Test
Sensibilidad (Recall)	0,7407 ± 0,1374	0,7425 ± 0,1437
Especificidad	0,7407 ± 0,1374	0,7425 ± 0,1437
Precisión	0,7554 ± 0,0758	0,7593 ± 0,0866
Valor predictivo negativo	0,7554 ± 0,0758	0,7593 ± 0,0866
Exactitud (accuracy)	0,7800 ± 0,0345	0,7826 ± 0,0493
F1-Score	0,7459 ± 0,1067	0,7486 ± 0,1138

Para la selección de características por el algoritmo Relief, hay una notable reducción de los valores en las medidas de rendimiento para entrenamiento y test en comparación con el análisis de componentes principales (Tabla 2).

Cabe destacar que, en ambos casos, la reducción sigue una tendencia similar, de tal forma que, atendiendo a las medidas de sensibilidad y especificidad en test, estas nos informan de un peor comportamiento de la red entrenada con las variables seleccionadas en esta aproximación para clasificar compuestos.

COMPARACIÓN DE MODELOS.

SELECCIÓN DE MEDIDA DE RENDIMIENTO PARA EL TEST ESTADÍSTICO.

Se realiza un test estadístico para comparar las aproximaciones de selección y extracción de características planteadas con las técnicas vistas en las prácticas anteriores. Para ello, los test se basan en las medidas del F1-Score en lo que respecta al test de las aproximaciones para reducir la dimensión, modelos de discriminante lineal, discriminante cuadrático y árboles de decisión.

RESULTADOS PARA LOS MODELOS IMPLICADOS. *ANÁLISIS DE COMPONENTES PRINCIPALES Y TÉCNICAS ANTERIORES*

Se muestran las medias y correspondientes desviaciones típicas para las medidas del F1-Score empleadas en el test estadístico para la comparación:

Tabla 3. Media y desviación típica de la medida de rendimiento F1-Score en test para comparar el análisis de componentes principales con las técnicas anteriores.

Modelo	F1-Score (media ± desviación típica)
Discriminante lineal	**0,8374 ± 0,0770**
Discriminante cuadrático	**0,7768 ± 0,0567**
Árbol de decisión	**0,7896 ± 0,0385**
SVM lineal	**0,8578 ± 0,0550**
Red neuronal	**0,8818 ± 0,0546**
Análisis de componentes principales	**0,8116 ± 0,1371**

Para el test se estableció un valor crítico de 0.10, que determina si las muestras correspondientes a las medidas de rendimiento en los modelos resultan o no similares estadísticamente. A continuación, se muestra la correspondiente tabla resultante del ANOVA de 1 vía:

Tabla 4. Tabla ANOVA de F1-Score en test para la comparación del análisis de componentes principales con las otras técnicas.

Source	SS	df	MS	F	Prob>F
Columns	0.01976	5	0.00395	3.77	0.0116
Error	0.02513	24	0.00105		
Total	0.04488	29			

El p-valor resultante es inferior al valor crítico de 0.10 establecido, de modo que se rechaza la hipótesis nula de ausencia de diferencias estadísticamente significativas entre la aproximación de análisis de componentes principales con el resto de las técnicas analizadas en las prácticas anteriores.

DIAGRAMA DE CAJAS Y BIGOTES.

La existencia de diferencias estadísticamente significativas se verifica posteriormente con la representación del diagrama de cajas y bigotes para el F1-Score sobre el conjunto de test como medida para la comparación de modelos:

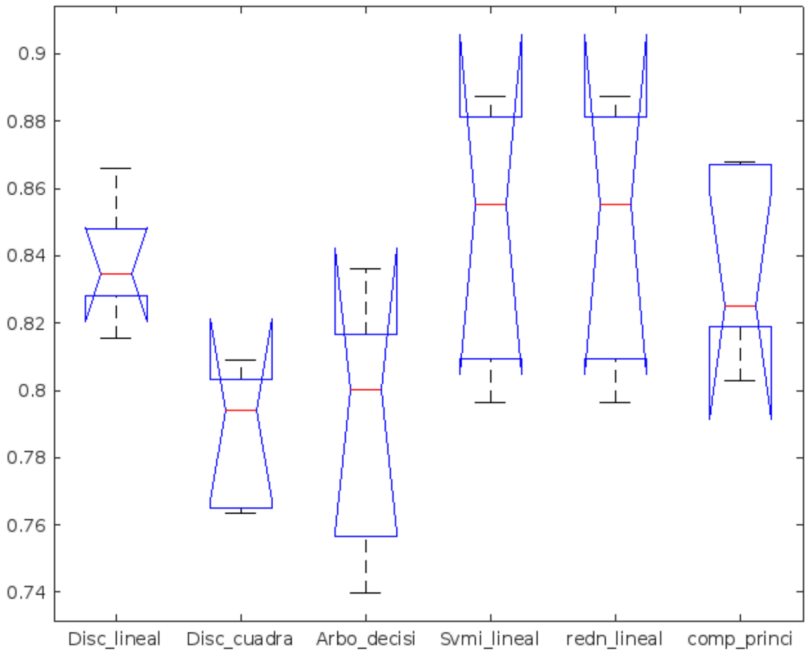

Figura 1. Diagrama de cajas y bigotes para F1-Score en test del análisis de componentes principales con las diferentes técnicas.

Del diagrama, se distingue principalmente que la red entrenada con las entradas transformadas presenta una menor mediana en comparación con la SVM lineal, la red neuronal de la práctica anterior (con mismos parámetros que la generada en la presente práctica) y el discriminante lineal. Por otra parte, cabe destacar que el rango intercuartílico es menor en comparación con la SVM lineal y la red neuronal.

TEST DE COMPARACIÓN MÚLTIPLE.

Dada la existencia de diferencias estadísticamente significativas entre las técnicas, se genera la representación del test de comparación múltiple, como un complemento al diagrama de cajas y bigotes:

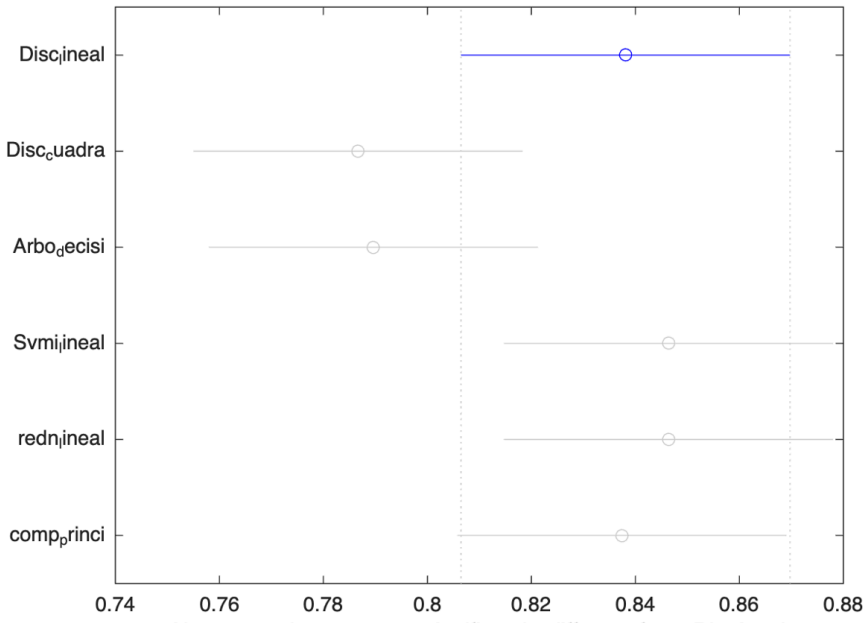

Figura 2. Representación del test de comparación múltiple para el análisis de componentes principales y las diferentes técnicas.

Este nos muestra la tendencia que se ha explicado anteriormente, distinguiendo aquellas técnicas entre las que se producen las mayores diferencias, que en este caso son el discriminante cuadrático junto con el árbol de decisión si se compara con el resto de las aproximaciones.

RESULTADOS PARA LOS MODELOS IMPLICADOS. *RELIEF Y TÉCNICAS ANTERIORES*

Se compara la técnica Relief para selección de características con los modelos de anteriores prácticas. A continuación, se muestran las medias y correspondientes desviaciones típicas para las medidas del F1-Score empleadas en el test estadístico para la comparación:

Tabla 5. Media y desviación típica de la medida de rendimiento F1-Score para comparación de Relief con técnicas anteriores.

Modelo	F1-Score (media ± desviación típica)
Discriminante lineal	0,8374 ± 0,0770
Discriminante cuadrático	0,7768 ± 0,0567
Árbol de decisión	0,7896 ± 0,0385
SVM lineal	0,8578 ± 0,0550
Red neuronal	0,8818 ± 0,0546
Relief	0,7486 ± 0,1138

De nuevo, para el test estadístico se estableció un valor crítico de 0.10. A continuación, se muestra la correspondiente tabla resultante del ANOVA de 1 vía:

Tabla 6. Tabla ANOVA de F1-Score en test para la comparación de Relief y técnicas vistas en las anteriores prácticas.

Source	SS	df	MS	F	Prob>F
Columns	0.04786	5	0.00957	6.3	0.007
Error	0.03649	24	0.00152		
Total	0.08436	29			

El p-valor resultante es inferior al valor crítico de 0.10 establecido, de tal forma que existen diferencias estadísticamente significativas entre la técnica de selección de características con Relief y el resto de las técnicas evaluadas en prácticas.

DIAGRAMA DE CAJAS Y BIGOTES.

Se verifica posteriormente la existencia de diferencias estadísticamente significativas con la representación del diagrama de cajas y bigotes para el F1-Score sobre el conjunto de test como medida para la comparación de modelos:

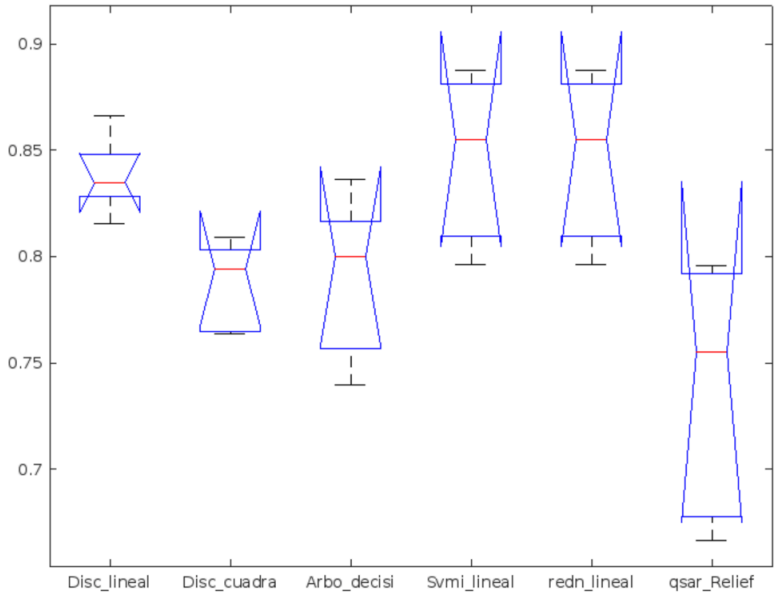

Figura 3. Diagrama de cajas y bigotes para F1-Score en Relief y el resto de las técnicas vistas.

Tal y como refleja en el diagrama obtenido, la red entrenada con las entradas transformadas presenta la mediana más baja en comparación con las otras aproximaciones, así como un notable rango intercuartílico. Hay una considerable diferencia si se compara el resultado con la SVM lineal y la red neuronal entrenada en la práctica anterior, que muestra los buenos resultados de estas aproximaciones frente a la técnica de selección de características tratada.

TEST DE COMPARACIÓN MÚLTIPLE.

Mediante la representación complementaria del test de comparación múltiple, se distinguen aquellas aproximaciones entre las que se producen diferencias estadísticamente significativas. En este caso, son más notables entre el modelo de discriminante lineal y la técnica de selección de características Relief:

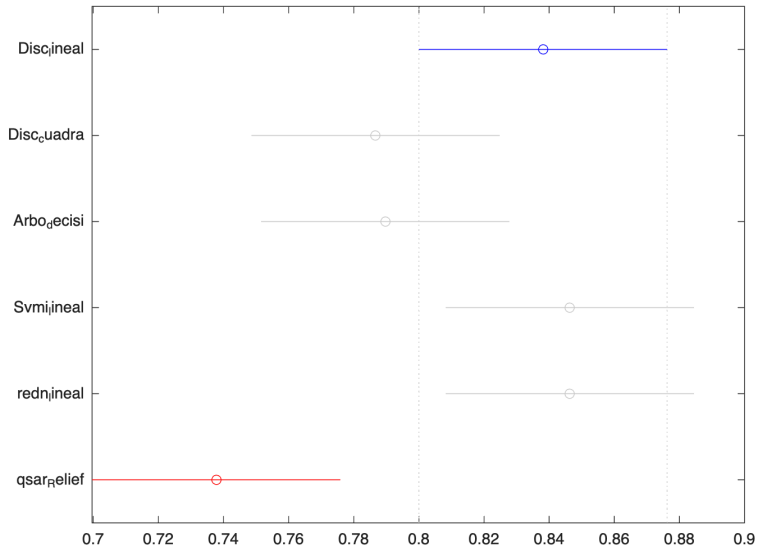

Figura 4. Representación del test de comparación múltiple para Relief y técnicas vistas.

DISCUSIÓN Y CONCLUSIONES.

Con las técnicas de reducción de dimensión empleadas en la presente práctica, se logran unos resultados interesantes para el conjunto de datos QSAR. En lo que respecta al análisis de componentes principales, con el mantenimiento de la varianza con un valor de 90 como mínimo, se alcanzan 22 componentes principales, lo que nos muestra una notable combinación lineal de las variables originales, para reducir correlación y explicar la mayor variabilidad existente en el conjunto de datos. El posterior entrenamiento de la red neuronal elegida con estas entradas transformadas muestra un peor resultado en la medida de rendimiento analizada. Esto podría asociarse a que la determinación de esa varianza implica una combinación lineal de variables originales que afecta a la relevancia para el conjunto tratado. No obstante, los resultados indican que el análisis de componentes principales muestra un mejor rendimiento si se compara con las técnicas del modelo de discriminante lineal y versión del árbol de decisión elegido en las anteriores prácticas.

Por otra parte, la selección de características por el método estadístico Relief muestra una tendencia similar, con una reducción más elevada en la medida de rendimiento si lo comparamos con el análisis de componentes principales. En base a esto, el número de características de 20 y el número de vecinos de 10 elegidos, muestran una influencia en la evaluación del algoritmo.

De esta forma, a continuación, se muestran las siguientes 10 características y pesos en cada una de ellas que nos proporciona MatLab:

Característica	25	39	13	30	38	1	3	36	37	24
Peso	0,0230	0,0211	0,0171	0,0124	0,0087	0,0083	0,0063	0,0062	0,0034	0,0014

Teniendo en cuenta estos resultados, junto con las 2 características que se emplean en la mitad (o más) de las ejecuciones realizadas, se determinaría la selección de las características 25 y 39 como aquellas que proporcionan una mayor distinción entre las clases del problema de QSAR.

A pesar de que la medida de rendimiento es peor en comparación con el resto de las técnicas vistas anteriormente, esta aproximación que se ha llevado a cabo permite adquirir una referencia en torno a posibles variables que conforman mayor relevancia para distinguir en el problema de clasificación. En concreto, estas características obtenidas corresponden con la presencia/ausencia de enlaces C-Cl a cierta distancia y con momentos espectrales.

Para futuros trabajos, sería interesante la realización de más pruebas en ambas aproximaciones, con el fin de evaluar aquellos parámetros que permitan lograr una mejora en la medida de rendimiento analizada y evaluar aquellas características con una mayor contribución en el conjunto de datos.