Experimentos computacionales en el dataset “ICTUS”

El dataset ICTUS está compuesto por 585 variables numéricas que describen a dos grupos de ictus. Hay un total de 18 muestras, 9 pertenecientes al grupo ICTUS-0 (N) y las otras 9 pertenecientes al grupo ICTUS-1 (S). El dataset no contiene valores perdidos.

En este estudio se realizaron una serie de experimentos computacionales con el objetivo de caracterizar los dos grupos de ictus, determinando conjuntos de proteínas que diferenciaran claramente a los dos grupos. Tener en cuenta que este problema cae dentro de los problemas conocidos en estadística como Fat-Short, al tener un número de variables descriptoras mucho mayor que número de muestras.

En la etapa de pre-procesamiento de datos se realizaron las siguientes operaciones:

1- Se comprobaron que no hubiera variables con varianza igual a cero o casi nula.
2- Se aplicó la transformación de Yeo-Johnson para eliminar la heterocedasticidad en los datos.
3- Se centraron y escalaron los datos.

A continuación se procedió a analizar el problema de la siguiente manera:

1- Se ejecutaron siete algoritmos de selección de variables (InfoGain, GainRatio, SymmetricalUncertainty, CFS, Consistency, ReliefF, Chi.squared) para estimar la importancia de las 585 proteínas. En todos los casos se utilizó un Leave-One-Out-Cross-Validation para estimar la importancia de las variables. Al final del proceso, la importancia de las proteínas se promediaron a lo largo de todas las ejecuciones y se generó un ranking, donde las importancias de las proteínas están en el rango [0,100]. El ranking general puede consultarse en la carpeta “feature-selection”.

Debido a la gran cantidad de proteínas, restringimos el análisis a aquellas proteínas que tuvieran una importancia mayor igual que 0.8 y además que sus niveles de expresión fueran significativamente diferentes entre los dos grupos de ictus. A partir de aquí el estudio se centró en las siguientes 21 proteínas.

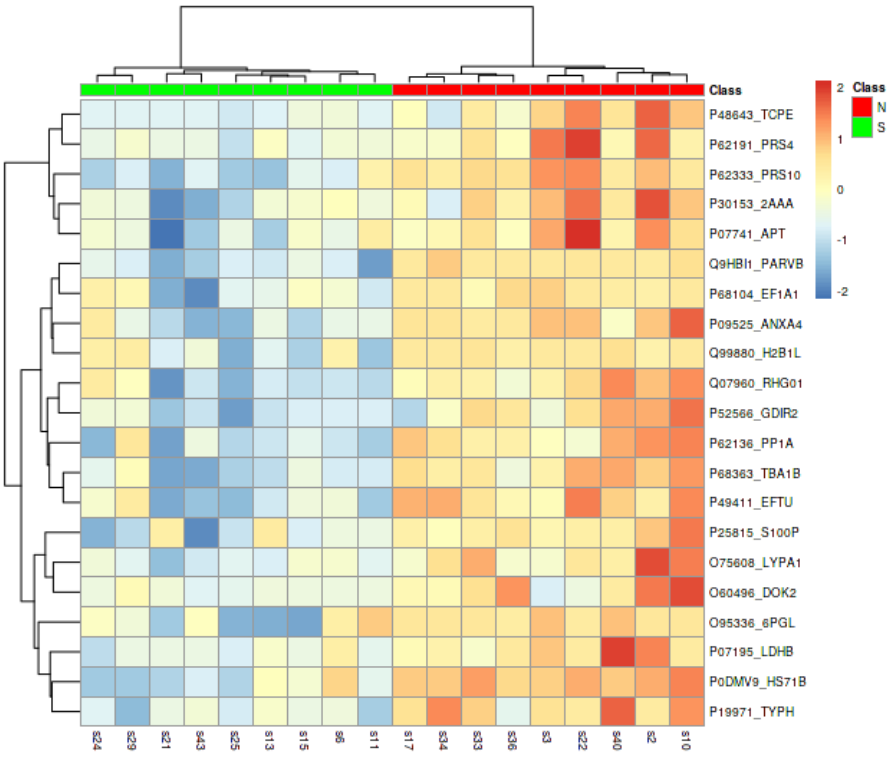
Var	Importance
Q9HBI1_PARVB	0.965
P62333_PRS10	0.853
P68363_TBA1B	0.853
P0DMV9_HS71B	0.841
P49411_EFTU	0.838
P62136_PP1A	0.837
P09525_ANXA4	0.836
P19971_TYPH	0.834
P07195_LDHB	0.833
P48643_TCPE	0.833
P68104_EF1A1	0.829
O95336_6PGL	0.825
P30153_2AAA	0.825
P62191_PRS4	0.818
Q99880_H2B1L	0.817
Q07960_RHG01	0.813
P52566_GDIR2	0.813

P07741_APT	0.809
P25815_S100P	0.807
O75608_LYPA1	0.805
O60496_DOK2	0.803

Las diferencias significativas entre los grupos de ictus pueden consultarse por los niveles de significación en la carpeta “Significant differences”.

2- Teniendo el ranking de 21 proteínas, se procedió a realizar un análisis de clustering haciendo una búsqueda heurística de combinaciones de proteínas que lograran un clustering perfecto de las muestras de ictus (la búsqueda heurística es la misma que se ha aplicado anteriormente en otros trabajos).

Se detectaron 398 combinaciones de proteínas que generan un clúster jerárquico con AUC igual a 1.0. Destacar que incluso considerando las 21 proteínas como un todo se logra un clustering con un AUC de 1.0, tal y como se muestra en la siguiente figura. Todos los gráficos heatmaps pueden consultarse en la carpeta “ranking-based-Clustering”.



3- Con el objetivo de filtrar más los resultados obtenidos, realizamos lo siguiente: (I) Consideramos dos algoritmos de clasificación bien conocidos en el ámbito de la biomedicina, Logistic Regression y Random Forest. (II) Filtramos aquellas combinaciones de proteínas que producen modelos con AUC igual a 1.0, tanto modelo de clustering, así como modelos de clasificación.

Con este procedimiento garantizamos que no solo se encuentre un buen modelo de clustering, sino que la combinación de proteínas también sea efectiva en la creación de modelos de clasificación, aumentando de esta manera la confiabilidad del conjunto de proteínas encontrado. Los modelos de

clasificación fueron estimados mediante un Leave-One-Out-Cross-Validation y se realizó un proceso de tunning para encontrar los mejores valores de sus parámetros.

Al final de este proceso se obtuvieron los siguientes 8 conjuntos de proteínas que cumplen las características antes descritas, es decir que generan tanto un clustering perfecto de los grupos de ictus, así como modelos perfectos de clasificación. Los heatmap de estas combinaciones de proteínas pueden consultarse en la carpeta “ranking-based-Classification-withFilteringperCluster”.

Conjuntos
1- Q9HBI1_PARVB P62333_PRS10
2- Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B
3- Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B P0DMV9_HS71B
4- Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B P0DMV9_HS71B P49411_EFTU
5- Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B P0DMV9_HS71B P49411_EFTU P62136_PP1A
6- Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B P0DMV9_HS71B P49411_EFTU P62136_PP1A P09525_ANXA4
7- Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B P0DMV9_HS71B P49411_EFTU P62136_PP1A P09525_ANXA4 P19971_TYPH
8- P62333_PRS10 P68363_TBA1B P0DMV9_HS71B P49411_EFTU P62136_PP1A P09525_ANXA4

Como puede observarse la mayoría de los conjuntos tienen como base el par de proteínas Q9HBI1_PARVB P62333_PRS10.

A partir de este análisis, podemos observar que el conjunto de las ocho proteínas Q9HBI1_PARVB P62333_PRS10 P68363_TBA1B P0DMV9_HS71B P49411_EFTU P62136_PP1A P09525_ANXA4 P19971_TYPH resulta significativo para diferenciar los dos grupos de ictus. Además, las muestras pertenecientes al grupo N (ictus=0 en el dataset original) se caracterizan por tener valores elevados en estas ocho proteínas, mientras que las muestras pertenecientes al grupo S (ictus=1) se caracterizan por tener valores bajos.

Por otra parte se calculó la correlación que existe entre estas ocho proteínas, así como entre las proteínas y la variable a predecir (aquella que representa los grupos de ictus). En todos los casos las correlaciones entre las ocho proteínas detectadas fueron altas, y por otro lado, se observa una correlación alta de las ocho proteínas con la variable que representa los grupos de ictus. Los niveles de correlación, así como los *p*-values asociados, pueden consultarse en la carpeta “Correlations”.