Entregable WEKA

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

Marzo 2020

Preprocesamiento

Después de analizar la base de datos, observamos que en ella no existen valores desconocidos. Como consecuencia, no tendremos que aplicar ninguna técnica de preprocesamiento para los valores desconocidos en esta base de datos. También observamos que la base de datos esta ordenada en función de las variables de la clase {Tumor, Normal}. En este caso, sí que tendremos que aplicar una técnica de preprocesamiento y procederemos a aleatorizar la base de datos. Para desordenar la base de datos aplicaremos un filtro a nivel de registro, concretamente de tipo no supervisado y de tipo registro llamado Ramdomize. Utilizaremos la semilla que viene por defecto.

A continuación, dividimos la base de datos en un conjunto de entrenamiento con dos tercios de los registros, y uno de test, con un tercio de los registros. Además lo haremos sin remplazo. Así, lo que tenemos que utilizar para ello, es un filtro de tipo no supervisado y de tipo registro llamado RemoveFolds. Los valores de este filtro que debemos modificar son: numFols = 3, fold = 1 y invertSelection = True. Como resultado hemos creado un conjunto de entrenamiento con 90 registros.

Clasificación

Selección de atributos

NaiveBayes

col	Accuracy	F-measure
NaiveBayes	cell5	cell6
C4.5	cell8	cell9

J48 (C4.5)

Técnicas de preprocesamiento

Selección de variables

AttributeSelectedClassifier

FilteredClassifier

Visualitzation

Confusion matrix ROC

1. inicialmente lanzaremos los clasificadores sobre la base de datos inicial, obteniendo así una medida inicial que intentaremos mejorar.

2. mejorar con la selección de variables

de variables es no lanzar una selección de tipo multivariada (filter o wrapper) sobre todas las variables, puesto que podría llevar mucho tiempo. Se recomienda primero hacer una selección univariada tipo ranker y quedarse con los mejores 200-500 atributos.