Entregable Discretización

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

Marzo 2020

El objetivo de esta práctica es analizar sobre la base de datos considerada, el efecto de incluir o no métodos de preprocesamiento, con los algoritmos de aprendizaje kNN y perceptrón multicapa, para mejorar las medidas iniciales de estos. Para ello se usa una validación cruzada de 5 carpetas (5cv) y la herramienta WEKA.

Exploración de los datos

Consideramos la base de datos vehicle definida sobre 18 variables predictivas (todas numéricas) y una variable clase multiclase {opel, saab, bus, van}. En ella no existen valores desconocidos, no está ordenada en función de la variable clase y está formada por 846 registros.

La distribución de la variable clase no es totalmente uniforme. Cuenta con:

- 212 registros para la etiqueta de la variable clase opel.
- 217 registros para la etiqueta de la variable clase saab.
- 218 registros para la etiqueta de la variable clase bus.
- 199 registros para la etiqueta de la variable clase van.

A continuación, se muestra la gráfica de la distribución de la variable clase.

GRÁFICA

Después usamos un filtro de tipo no supervisado y de registro, llamado RemoveFolds proporcionado por la herramienta WEKA, que elimina alguna de las carpetas en una validación cruzada, para crear un conjunto de entrenamiento que contendrá dos tercios de los registros de la base de datos. Una vez aplicado, el número de registros se reduce a 564.

La distribución de la variable clase sigue siendo no uniforme y ahora cuenta con:

- 149 registros para la etiqueta de la variable clase opel.
- 153 registros para la etiqueta de la variable clase saab.
- 139 registros para la etiqueta de la variable clase bus.

- 123 registros para la etiqueta de la variable clase van.

A continuación, se muestra la gráfica de la distribución de la variable clase del conjunto de entrenamiento.

GRÁFICA

Algoritmos de preprocesamiento

Los algoritmos que se describen a continuación son:

-
-
-

Resultados y análisis

Se han considerado dos parámetros de rendimiento para la evaluación de los resultados. Los siguientes parámetros son examinados tanto antes como después de la discretización: Accuracy y Error Rate.

La siguiente tabla proporciona la precisión y la tasa de error para cada clasificador antes y después de la discretización.

Clasificador	Antes Disc.		Después Disc.	
	Acc in %	ERR in %	Acc in %	ERR in %
kNN (k=1)	64.7163	35.2837	2.1	2.1
kNN (k=3)	66.844	33.156	2.1	2.1
Multilayer Perceptron	79.4326	20.5674	11.6	11.6