REDUCCIÓN DE LA DIMENSIONALIDAD UTILIZANDO SELECCIÓN DE INSTANCIAS EN DATOS MULTI-ETIQUETA

Adis perla Mariño Rivero[1] y Sebastián Ventura Soto[1]

1 Universidad de Córdoba. mrivero@yandex.com

El Descubrimiento de Conocimiento en Bases de Datos es un área de la computación que intenta analizar grandes volúmenes de datos, extrayendo conocimiento útil que pueda asistir a un humano para llevar a cabo tareas de forma más eficiente y satisfactoria. Debido al tamaño de las bases de datos, la presencia de ruido, datos inconsistentes y redundantes, se hace necesaria la aplicación de técnicas de preprocesamiento sobre los conjuntos de datos. La Selección de Instancias (IS) es una importante área dentro del preprocesamiento de datos que tiene como principal objetivo la mejora de la generalización de los modelos de predicción a partir de la selección de los ejemplos más representativos. El conocido problema “maldición de la dimensionalidad” también está presente a la hora de aprender modelos a partir de datos multi-etiqueta. Hasta la fecha se reportan muy pocos trabajos en la literatura utilizando técnicas de IS sobre este tipo de datos. En este trabajo se hace un estudio empírico de los métodos tradicionales de IS que combinados con técnicas de transformación de problemas se aplican a conjuntos de datos multi-etiqueta de diferentes dominios como texto, biología, música, imágenes y videos. Los resultados obtenidos a partir del estudio realizado demuestran que los algoritmos de aprendizaje automático obtienen mejores resultados sobre los conjuntos de datos reducidos, manteniendo niveles de precisión significativamente iguales o superiores a cuando el aprendizaje se realiza sobre los conjuntos de datos originales.