UNIVERSIDAD DE GRANADA

☐

Línea de Trabajo fin de Máster

(Fecha última actualización: 27/09/22)

Máster en Estadística Aplicada. CURSO ACADÉMICO 2022-2023	
Título	Análisis de la eficiencia de técnicas de filtrado de ruido en problemas de clasificación con datos genómicos
Tipo	INVESTIGACIÓN ☒ ORIENTACIÓN PRÁCTICA ☐
Profesor(es)/ email	Pedro María Carmona Sáez (pcarmona@ugr.es), José Antonio Sáez Muñoz (joseasaezm@ugr.es)
Descripción	La adquisición y el procesamiento de datos en aplicaciones estadísticas y de minería de datos a menudo están sujetos a imperfecciones. Este hecho puede dar lugar a la presencia de errores o ruido en los conjuntos de datos. En problemas de clasificación, la creación de modelos a partir de datos ruidosos presenta varios inconvenientes, entre los que están la necesidad de más tiempo y muestras para construir el clasificador. Además, tanto la precisión como la complejidad de los clasificadores pueden verse afectadas por el modelado de datos corruptos. En este contexto, los filtros de ruido se han propuesto como una alternativa para paliar las consecuencias negativas que tienen los errores en los datos. Estos se basan en eliminar muestras con ruido de clase del conjunto de datos de entrenamiento. La separación de la detección de ruido y el aprendizaje tiene la ventaja de que las muestras ruidosas no influyen en la construcción del modelo. Así, la eliminación de dichas muestras reduce el tamaño del conjunto de datos original mediante la selección de datos relevantes, lo que mejora el rendimiento de los modelos aprendidos posteriormente. Pese a su relevancia, este tipo de técnicas usualmente no se emplea en el campo de la clasificación con datos genómicos, donde el ruido de etiqueta puede ser común debido a errores humanos en el etiquetado. El trabajo realizado pretende analizar la aplicación de las técnicas de filtrado de ruido en problemas genómicos con el objetivo de estudiar las sinergias existentes al tratar con este tipo de datos, caracterizados particularmente por una alta dimensionalidad.
Objetivos particulares	- Desarrollar tareas de investigación en un entorno multidisciplinar - Entender algunos de los principales retos actuales en la investigación en el campo - Aplicar técnicas de preprocesamiento de datos y clasificación en datos genómicos
Prerrequisitos y recomendaciones	• Conocimientos avanzado de lenguaje R. • Conocimiento de técnicas de clasificación
Plan de trabajo	- Revisión bibliográfica - Análisis comparativo de las diferentes metodologías y paquetes de R - Implementación de funciones en R para filtrado y clasificación - Aplicación a un conjunto de datos reales - Interpretación de resultados
Competencias generales y específicas	CB: 6, 7, 8, 9, 10 CG: 1, 2, 3, 6, 9 CE: 10, 13, 15, 16, 17, 18, 19, 20, 22, 23, 26, 28, 29
Bibliografía	• C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data", J. Artif. Intell. Res., vol. 11, pp. 131–167, Aug. 1999. • T.M.Khoshgoftaar and P.Rebours, "Improving software quality prediction by noise filtering techniques," J. Comput. Sci. Technol., vol. 22, no. 3, pp. 387–396, May 2007. • Nematzadeh, Z.; Ibrahim, R.; Selamat, A. Improving class noise detection and classification performance: A new two-filter CNDC model. Appl. Soft Comput. 2020, 94, 106428. • Liu, T.; Tao, D. Classification with noisy labels by importance reweighting. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 447–461.