SMOTE-BD: an Exact and Scalable Oversampling Method for Imbalanced Classification in Big Data

María José Basgall, Waldo Hasperué, Marcelo Naiouf, y otros

Journal of computer science and technology, ISSN-e 1666-6038, Vol. 18, n. 3, 2018, p. 203-209

El volumen de datos en las aplicaciones de hoy en día ha significado un cambio en la forma de abordar los problemas de Machine Learning. De hecho, el escenario Big Data implica restricciones de escalabilidad que sólo se pueden lograr a través del diseño de modelos inteligentes y el uso de tecnologías distribuidas. En este contexto, las soluciones basadas en la plataforma Spark se han establecido como un estándar de facto. En esta contribución, nos centramos en un marco muy importante dentro de Big Data Analytics, a saber, la clasificación con conjuntos de datos desequilibrados. La principal característica de este problema es que una de las clases está sub-representada y, por lo tanto, generalmente es más complejo encontrar un modelo que la identifique correctamente. Por esta razón, es común aplicar técnicas de preprocesamiento como el sobremuestreo, para equilibrar la distribución de ejemplos en las clases. En este trabajo presentamos SMOTE-BD, un enfoque de preprocesamiento totalmente escalable para la clasificación no balanceada en Big Data. El mismo se basa en una de las soluciones de preprocesamiento más extendidas para la clasificación desequilibrada, a saber, el algoritmo SMOTE, el cual crea nuevas instancias sintéticas de acuerdo con la vecindad de

cada ejemplo de la clase minoritaria. Nuestro novedoso desarrollo está hecho para ser independiente de la cantidad de particiones o procesos creados, para lograr un mayor grado de eficiencia. Los experimentos realizados en diferentes conjuntos de datos estándar y de Big Data muestran la calidad del diseño y la implementación propuestos.

Resumen realizado por la propia publicación