Generación de datos artificiales para clases no balanceadas en Big Data

Generation of artificial data for imbalanced classes in Big Data

MSc. Yordan Ernesto Estrada Rodríguez[a]*, Dr. Luis Carlos Méndez González[a]

[a]Departamento de Ingeniería Industrial y Manufactura, Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, Chihuahua, 32310, México

*Autor de correspondencia. Correo: al216923@alumnos.uacj.mx

No. de resumen
2CP21-33

Formato
Cartel

Evento
2.º Coloquio de Posgrados IIT

Presentador
Yordan Ernesto Estrada Rodríguez

Tema
Procesos Tecnológicos

Estatus
Estudio en curso

Fecha de la presentación
Noviembre 11, 2021

RESUMEN

Los algoritmos de *machine learning* han surgido en la actualidad como una solución viable a problemas de la industria y diversas áreas de las ciencias. Sin embargo, su entrenamiento constituye a la vez su principal fortaleza y debilidad. Una máquina capaz de aprender y tomar decisiones constituye una herramienta única aplicable, incluso en ambientes nocivos para el ser humano. En este aspecto, la información contenida en el ámbito de Big Data se traduce en conocimiento para el entrenamiento de los mismos. Esta investigación propone el desarrollo de una metodología capaz de solucionar el problema de clases desbalanceadas en presencia de altos volúmenes de datos. Para ello se pretende identificar el estado actual de las herramientas Big Data en términos de rendimiento, así como el estado del arte de las técnicas de generación de datos artificiales. Como resultado previo, es posible puntualizar que en el ámbito de Big Data, la mayoría de las herramientas actuales no son capaces de proveer resultados satisfactorios, pues la presencia de altos volúmenes de información conlleva a nuevos retos, donde la cantidad de datos y la escalabilidad de los algoritmos constituyen verdaderos desafíos. Tomando en consideración la cantidad de datos que plantea el uso de Big Data, donde la cantidad de elementos facilita la identificación de distribuciones estadísticas en los subconjuntos de entrenamientos para clasificadores. En este estudio se propone un algoritmo de sobremuestreo distribuido, centrado en densidades estadísticas, capaz de tener en cuenta el comportamiento de los valores de cada atributo y, por tanto, de respetar su distribución.

Palabras clave: datos artificiales, aprendizaje automático, distribuciones estadísticas, Big Data.

ABSTRACT

Machine learning algorithms have emerged today as a viable solution to problems in industry and various areas of science. However, its training is both its main strength and weakness. A machine capable of learning and making decisions constitutes a unique tool applicable even in environments that are harmful to humans. In this aspect, the information contained in the field of big data translates into knowledge for their training. This research proposes the development of a methodology capable of solving the problem of unbalanced classes in the presence of high volumes of data. To do this, it is intended to identify the current state of big data tools in terms of performance, as well as the state of the art of artificial data generation techniques. As a previous result, it is possible to point out that in the field of Big Data, most of the current tools are not capable of providing satisfactory results, since the presence of high volumes of information leads to new challenges, where the amount of data and the scalability of algorithms are real challenges. Taking into consideration that the amount of data represented by the use of big data facilitates the identification of statistical distributions in the training subsets of classifiers, a distributed oversampling algorithm is proposed, focused on statistical densities and capable of having in it takes into account the behavior of the values of each attribute respecting its distribution.

Keywords: artificial data, machine learning, statistical distributions, Big Data.

Entidad legal responsable del estudio

Universidad Autónoma de Ciudad Juárez.

Financiamiento

El financiamiento de esta investigación se lleva gracias al apoyo del sistema de Becas de Posgrado Nacionales de CONACYT, clave 2021-000018-02NACF.

Conflictos de interés

Los autores declaran que no existe conflicto de intereses.