Algoritmo para la generación de datos mediante la distribución de probabilidad de los atributos en clases no balanceadas de Big Data

Algorithm for the generation of data through the probability distribution of the attributes in unbalanced classes of Big Data

Yordan Ernesto Estrada Rodríguez[a]*, Luis Carlos Méndez González[a]
[a]Doctorado en Tecnología, Departamento de Ingeniería Industrial y Manufactura, Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, México.
*Autor de correspondencia. Correo electrónico: al216923@alumnos.uacj.mx

No. de resumen
4CP22-41

Evento
4.º Coloquio de Posgrados del IIT

Tema
Cómputo Aplicado

Fecha de la presentación
Noviembre 25, 2022

Formato
Ponencia

Presentador
Yordan Ernesto Estrada Rodríguez

Estatus
Estudio en curso

Resumen

El continuo crecimiento de la información, el auge de la analítica de datos y la tendencia cada vez más frecuente a la automatización de procesos mediante el uso de algoritmos de Inteligencia artificial, demandan metodologías que posibiliten el entrenamiento de algoritmos más eficientes y con menor tendencia al error. La información contenida en los grandes volúmenes de datos actuales constituye una opción viable para el entrenamiento de los algoritmos de Machine Learning. La presente investigación propone una técnica híbrida y escalable capaz de resolver el problema de las clases no balanceadas en un ambiente con grandes volúmenes de información. La misma posibilitará hacer uso de las bondades del Big Data para la generación de conjuntos de datos que tributarán a un correcto entrenamiento de los algoritmos de Machine Learning. El desarrollo positivo de la investigación propuesta supone el desarrollo de un algoritmo capaz de generar datos según el comportamiento de los datos, lo cual incidiría directamente en las ramas de la industria y las ciencias que hagan uso de estos algoritmos para la toma de decisiones, optimizando los procesos y logrando resultados económicos y sociales. Esta investigación se puede englobar dentro del contexto de Big Data Analytics, la cual es una áreas más importantes y rentables dentro de la ciencia de datos.

Palabras clave: datos desbalanceados, generación de datos, distribuciones estadísticas, Big Data.

Abstract

The continuous growth of information, the rise of data analytics and the increasingly frequent trend towards process automation with artificial intelligence algorithms, demand methodologies that enable the training of more efficient algorithms with less tendency to mistake. The information contained in today's large volumes of data is a viable option for training Machine Learning algorithms. This research proposes a hybrid and scalable technique capable of solving the problem of unbalanced classes in an environment with large volumes of information. It will make it possible to make use of the benefits of Big Data for the generation of data sets that will contribute to a correct training of Machine Learning algorithms. The positive development of the proposed research supposes the development of an algorithm capable of generating data according to the behavior of the data, which would directly affect the branches of industry and sciences that make use of these algorithms for decision making, optimizing

processes and achieving economic and social results. This research can be included within the context of Big Data Analytics, which is one of the most important and profitable areas within data science.

Keywords: unbalanced data, data generation, statistical distributions, big data.

Entidad legal responsable del estudio
Universidad Autónoma de Ciudad Juárez.

Financiamiento
Universidad Autónoma de Ciudad Juárez - beca posgrado CONACYT.

Conflictos de interés
Los autores declaran que no existe conflicto de intereses.