Clases no balanceadas y alta dimensionalidad en Big Data: Un estudio de SMOTE con normas fraccionales

Class imbalance and high dimensional Big Data: A study of SMOTE with fractional norms

Armando Isaac Bolívar Velazco[a], Vicente García Jiménez[a]*, Rogelio Florencia Juárez[a], Roberto Alejo Eleuterio[b]
[a]Departamento de Ingeniería Eléctrica y Computación, Doctorado en Ciencias de la Ingeniería Avanzada, Universidad Autónoma de Ciudad Juárez, México
[b]Instituto Tecnológico de Toluca, Metepec, Estado de México, México

No. de resumen
CIP21-46

Evento
1.er Coloquio de Investigación y Posgrado

Tema
Cómputo Aplicado (COAP)

Fecha de la presentación
21 de mayo de 2021

Formato
Ponencia

Presentador
Primer/a autor/a en la lista de autores

Estatus
Resultados preliminares

RESUMEN

La presente investigación tiene como objetivo analizar el comportamiento de un algoritmo de sobremuestreo basado en distancias, llamado SMOTE, en un problema de clases no balanceadas y alta dimensionalidad en Big Data. Para ello, se obtuvo una base de datos del repositorio LIBSVM, en específico se trabajó con el conjunto de datos de KDD 2010 que cuenta con 2 clases, 19 264 097 instancias y 1 163 024 atributos. La base de datos es dispersa, por lo que algunos atributos tienen un valor de 0. Para la experimentación se utilizaron tres máquinas virtuales en la nube de Azure, una instancia maestra/ejecutora Standard_E2as_v4 y dos instancias ejecutoras Standard_D2ds_v4 configuradas en clúster con Spark 3.1.1. Debido a las limitantes del equipo de cómputo para acomodar todos los datos en memoria, se redujo la cantidad de instancias a 30 000. Para evaluar el efecto de la alta dimensionalidad en el algoritmo de sobremuestreo, la base de datos fue modificada con diferentes tamaños que van desde 50 hasta 900 atributos. En todos los casos, el radio de desbalance fue de 1:10. Asimismo, se llevaron a cabo experimentos usando normas fraccionales en el SMOTE, que pueden ayudar a reducir el efecto de la alta dimensionalidad. Los resultados obtenidos usando un árbol de decisión muestran que existe una mejora en la clasificación de la clase minoritaria cuando se emplea SMOTE con normas fraccionales. Sin embargo, se observó que la tasa de reconocimiento en la clase mayoritaria se vio drásticamente reducida. Esto quizás se deba que el conjunto de datos es disperso, por lo que puede haber ocasionado un sobreajuste del clasificador. En la siguiente etapa de esta investigación se realizará la misma experimentación, pero empleando un conjunto de datos no disperso.

Palabras clave: SMOTE; Big Data; alta dimensionalidad; clases no balanceadas; normas fraccionales.

ABSTRACT

The present research aims to analyze the behavior of an oversampling algorithm, called SMOTE, on an unbalanced and high-dimensional Big Data set. Here we employed the KDD 2010 dataset obtained from the LIBSVM repository. Some features have zero values, so the dataset is considered sparse. The experiments were executed in cloud-based clusters provided by the Azure Platform. The cloud cluster was composed of three VMs, one master/executor Standard_E2as_v4, and two Standard_D2ds_v4 nodes with Spark 3.1.1. Due to the huge size of the datasets and computer limitations, the number of instances was reduced to 30,000. Besides, several datasets were created with different dimensionality sizes ranging from 50 to 900 attributes. In all cases, the imbalanced ratio was 1:10. To reduce the effect of high dimensionality, the datasets were preprocessed with SMOTE and fractional distances. Results obtained using a decision tree show an improvement in minority class

classification when using SMOTE with fractional norms. However, the recognition rate in the majority class was drastically reduced. As the dataset is sparse, we believe that it may derive into an overfitted model. The next stage of this research will perform the same experiment but using a non-sparse dataset.

Keywords: SMOTE; Big Data; high dimensionality; class imbalance; fractional norms.

*Autor de correspondencia. Correo electrónico: vicente.jimenez@uacj.mx