Estudio experimental de técnicas de sobremuestreo en conjuntos de datos masivos desbalanceados con baja y alta densidad

Experimental study of oversampling techniques for imbalanced big data sets with sparse and non-sparse representation

Armando Isaac Bolívar Velazco[a], Vicente García Jiménez[a]*, Rogelio Florencia Juárez[a], Roberto Alejo Eleuterio[b]

[a]Departamento de Eléctrica y Computación, Universidad Autónoma de Ciudad Juárez, México
[b]Instituto Tecnológico de Toluca, Metepec, Estado de México, México
*Autor de correspondencia. Correo: vicente.jimenez@uacj.mx

No. de resumen
2CP21-26

Formato
Ponencia

Evento
2.º Coloquio de Posgrados IIT

Presentador
Armando Isaac Bolívar Velazco

Tema
Cómputo Aplicado

Estatus
Estudio en curso

Fecha de la presentación
Noviembre 12, 2021

RESUMEN

En este trabajo se analiza el comportamiento de dos técnicas de sobremuestreo enfocadas en tratar el problema de clasificación de datos masivos desbalanceados, cuando adicionalmente se presenta una alta dimensionalidad. La literatura menciona que, en conjuntos de datos con un gran número de atributos, las técnicas de sobremuestreo basadas en distancias euclidianas se ven afectadas. Se obtuvo una base de datos denominada KDD 2010 del repositorio LIBSVM, que cuenta con dos clases, 19 264 097 instancias y 1 163 024 dimensiones. La base de datos es de baja densidad, por lo que la mayoría de los atributos contienen ceros. Para generar una base de datos con alta densidad se empleó un PCA. Los experimentos se realizaron en la nube pública de Google, donde se configuro un clúster de Spark 3.1.2 con un nodo maestro y cuatro nodos esclavos. Como algoritmos de sobremuestreo y clasificación se usaron ROS, SMOTE, SVC y árbol de decisión. En la base de datos de baja densidad, el fenómeno de la maldición de la dimensionalidad no parece afectar de manera evidente el cálculo de distancias de SMOTE, sino que, paradójicamente, entre mayor la dimensionalidad mejor es la tasa de clasificación. Caso contrario se observa en la base de datos con alta densidad, donde conforme se incrementan las dimensiones se observa un deterioro de la eficacia de SMOTE. Los efectos de la maldición de la dimensionalidad se podrían definir en términos de el número de atributos y la densidad. SMOTE no se ve afectado en conjuntos de datos con alta dimensionalidad y baja densidad.

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

ABSTRACT

In this work, the behavior of two oversampling techniques focused on treating the classification problem of massive unbalanced data is analyzed when high dimensionality is presented. The literature mentions that in data sets with a large number of attributes, oversampling techniques based on Euclidean distances are affected. A database called KDD 2010 was obtained from the LIBSVM repository that has two classes, 19 264 097 instances, and 1 163 024 dimensions. The database is sparse, so most attributes contain zeros. To generate a dense database, a PCA was used. The experiments were conducted on Google's public cloud. A Spark 3.1.2 cluster was configured with one master node and four slave nodes. ROS, SMOTE, SVC, and decision tree were used as oversampling and classification algorithms. It was observed that in the sparse database, the phenomenon of the curse of dimensionality does not seem to affect the calculation of SMOTE distances; but, paradoxically, the higher the dimensionality, the better the classification rate. The opposite case is observed in the dense database where, as the dimensions are increased, a deterioration in the effectiveness of SMOTE is observed. The effects of the curse of dimensionality could be defined in terms of the number of attributes and density. SMOTE is affected only in high dimensionality and non-sparse datasets.

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

Entidad legal responsable del estudio

Universidad Autónoma de Ciudad Juárez.

Financiamiento

Ninguno.

Conflictos de interés

Los autores declaran que no existe conflicto de intereses.