Técnicas de sobremuestreo *big data* en espacios de disimilitud en conjuntos de datos desbalanceados con alta dimensionalidad y solapamiento

Big data oversampling techniques in dissimilarity spaces on imbalanced datasets with high dimensionality and overlapping

Armando Isaac Bolívar Velazco[a], Vicente García Jiménez[a]*, Rogelio Florencia Juárez[a], Roberto Alejo Eleuterio[b]
[a]Departamento de Ingeniería Eléctrica y Computación, Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, México.
[b]Instituto Tecnológico de Toluca, Metepec, Estado de México, México.
*Autor de correspondencia. Correo electrónico: vicente.jimenez@uacj.mx

No. de resumen
4CP22-23

Formato
Ponencia

Evento
4.º Coloquio de Posgrados del IIT

Presentador
Armando Isaac Bolívar Velazco

Tema
Cómputo Aplicado

Estatus
Estudio en curso

Fecha de la presentación
Noviembre 24, 2022

Resumen

En este trabajo se usa una técnica de transformación del espacio llamada disimilitud para mitigar el problema del solapamiento y la alta dimensionalidad. Además, se utiliza SMOTE con normas fraccionarias y con distancia euclidiana en el espacio original y de disimilitud. También, se caracteriza el solapamiento a nivel de atributos utilizando la métrica de complejidad F1. Para ello, se generaron bases de datos sintéticas solapadas y desbalanceadas con una relación de desbalance de 10:1 (mayoritaria: minoritaria), 110 000 ejemplos, dimensionalidad que va de 1000 hasta 4000 dimensiones y un 20 % de ruido. Los experimentos se realizaron en la nube de Google, donde se configuró un clúster de Spark 3.1.2 con un nodo maestro y siete nodos esclavos. Se compararon los resultados obtenidos de SMOTE con normas fraccionarias en el espacio de características y en el espacio de disimilitud. SMOTE con normas fraccionarias en el espacio de características obtuvo el mejor desempeño de TPR, mientras que SMOTE en el espacio de disimilitud obtuvo el mejor AUC-ROC. Cuando se comparó el solapamiento de características por medio de la métrica F1, en el nuevo espacio se logró disminuir el solapamiento. En trabajos futuros, se buscará también tratar el solapamiento a nivel de instancias por medio de técnicas basadas en el vecino más cercano.

Palabras clave: SMOTE, *big data*, alta dimensionalidad, clases no balanceadas, disimilitud.

Abstract

In this work, a space transformation technique called dissimilarity is used to mitigate the problem of overlapping and high dimensionality. In addition, SMOTE is used with fractional norms and Euclidean distance in the original and dissimilarity space. Also, the overlap is characterized at the attribute level using the F1 complexity metric. For this, overlapping and unbalanced synthetic databases were generated with an imbalance ratio of 10:1 (majority: minority), 110,000 examples, dimensionality ranging from 1,000 to 4,000 dimensions, and 20% noise. The experiments were performed in the Google Cloud, where a Spark 3.1.2 cluster was configured with one master node and seven slave nodes. The results obtained from SMOTE were compared with fractional norms in the feature and dissimilarity spaces. SMOTE with fractional norms in the feature space obtained the best performance from TPR and, in con-

trast, SMOTE in the dissimilarity space obtained the best AUC-ROC. When the overlap of characteristics was compared using the F1 metric, the overlap was reduced in the new space. In future works, we will also seek to deal the overlap at the instance level through techniques based on the nearest neighbor.

Keywords: SMOTE, big data, high dimensionality, class imbalance, dissimilarity.

Entidad legal responsable del estudio

Universidad Autónoma de Ciudad Juárez.

Financiamiento

Créditos para investigación de Google Cloud.

Conflictos de interés

Los autores declaran que no existe ningún conflicto de interés.