RJIT Research Journal of Information Technology

OPEN ACCESS

Executive Summary

# A Specialized Synthetic Minority-Oversampling Method

*Asif Ali*

A classification issue with an uneven class distribution is imbalanced data. Human mistake, the lack of samples relevant to a particular class, or other factors can cause the distribution of the class samples to be uneven. This group is frequently referred to as the minority class. In other words, there are fewer items in the minority or positive class than in the majority or negative class. Researchers and academicians may find it difficult to improve a classification model's performance while using all available machine learning techniques. As a result, this class imbalance leads to several difficult problems, including incorrect data element distribution, class overlap, a class with noise, the sample size of training data, etc.

To correct the erroneous data distribution, researchers have suggested several strategies for rebalancing the data samples within minority and majority classes. Outlier-SMOTE, a different modified form of SMOTE in which the outliers are determined using Euclidian distance, has been introduced. The main goal of this research was to improve the minority class elements and create a hybrid model for a new generation of synthetic data elements. The suggested methodology cannot be validated by merely measuring training and testing accuracy. The study used an imbalanced version of the Pima dataset with two classes, positive and negative, with no missing values, as well as five datasets from the Keel dataset repository to assess the performance of the suggested model.

The goal of this study was to develop the hybrid meta-heuristic model PSO-EV for data augmentation to address problems with data imbalance caused by datasets with an unbalanced distribution of data items among their classes. To improve classification performance, an effort has been made to collect the optimum synthetic samples by PSO and EV and supplement those freshly created synthetic samples towards the minority class centroid. In this study, the system is first given the unbalanced dataset inputs, and then SMOTE is used to produce artificial samples. The updated velocity and position of the data elements are then obtained by creating a set of optimized synthetic samples using PSO. Finally, the experimentation and result analysis show that the suggested SMOTE-PSOEV performs admirably for all five datasets used for experimentation.

This work explores the properties and capabilities of two meta-heuristic optimization algorithms and introduces a novel variant of SMOTE called SMOTE-PSOEV. The suggested methodology combines SMOTE to create synthetic samples initially, and PSO and EV to improve those samples. SMOTE-PSOEV experimentation and validation have been conducted over the course of three phases. Three classifiers, including SVM, NB, and k-NN, have their recognition rates reported. Finally, the

experimental results demonstrate that SMOTE-PSOEV outperforms other SMOTE versions and can mine the data for those investigated datasets despite imbalanced class distribution.

Source: Information

KEYWORDS
Class imbalance problem; data augmentation; SMOTE; particle swarm optimization; Egyptian vulture