Under-Sampling the Minority Class to Improve the Performance of Over-Sampling Algorithms in Imbalanced Data Sets

Romero F. A. B. de Morais (rfabm@cin.ufpe.br)
Germano C. Vasconcelos (gcv@cin.ufpe.br)

Center for Informatics
Federal University of Pernambuco

IJCAI Workshop on Learning in the Presence of Class Imbalance and Concept Drift, August 2017
Table of Contents

Motivation

Neighbourhood of Influence

$k$-INOS

Experimentation

Summary
Table of Contents

Motivation

Neighbourhood of Influence

$k$-INOS

Experimentation

Summary
Motivation

- Many over-sampling algorithms available.

- Majority of them utilise all the examples in the minority class during the over-sampling process.

- ADASYN, SMOTE, RWO, …

- Under-sampling the minority class before over-sampling is rarely attempted.
Table of Contents

Motivation

Neighbourhood of Influence

\( k \)-INOS

Experimentation

Summary
Imbalanced Scenario
Imbalanced Scenario
$k$-Nearest Neighbours
Reverse $k$-Nearest Neighbours
\textit{k-Influential Neighbourhood}
Modified $k$-Influential Neighbourhood
Table of Contents

Motivation

Neighbourhood of Influence

$k$-INOS

Experimentation

Summary
**$k$-INOS Algorithm**

**input**: D: Imbalanced Data Set  
  k: Number of neighbours to compute $k$-IN  
  $\tau$: $k$-IN size threshold  
  $\phi$: Over-sampling function  

**output**: D*: A more balanced version of D

1. For each minority class example in D compute its modified $k$-IN  
2. Remove from D all the minority class examples that have a modified $k$-IN smaller than $\tau$  
3. Call $\phi$ on D  
4. Add back the examples removed in the second step to the over-sampled data
2-D Example - SMOTE vs. $k$-INOS
Mammography Data Set
Mammography Data Set - SMOTE vs. $k$-INOS
Table of Contents

Motivation

Neighbourhood of Influence

k-INOS

Experimentation

Summary
Settings

- 50 imbalanced data sets.
- 5 base classifiers.
- 7 over-sampling algorithms.
- 5 performance metrics.
- $5 \times 2$-fold cross-validation to assess performance.
- Wilcoxon signed-ranks test to analyse performance difference between over-sampling algorithms with and without $k$-INOS.
Results

Accuracy  Significantly increased for most combinations of classifier and over-sampling algorithm.

AUROC   Increased most of the time for the GBM and 3-NN classifiers and half the time for DT.

F1     Increased most of the time for the DT, GBM, 3-NN, and SVM classifiers. Many significant increases for the DT, GBM, and 3-NN classifiers.

Recall Significantly decreased for most combinations of classifier and over-sampling algorithm.

Precision Significantly increased for almost all combinations of classifier and over-sampling algorithm.
Table of Contents

Motivation

Neighbourhood of Influence

$k$-INOS

Experimentation

Summary
Advantages

- A general wrapper for over-sampling algorithms.
- Increases the performance of most metrics especially for weak classifiers.
- Easy to implement.
Disadvantages

▶ Computation of the neighbourhood of influence might be expensive.

▶ Does not seem to work well with strong classifiers.

▶ Decreases Recall.
Future Work

- Analyse in which situations $k$-INOS is likely to attain performance improvements.

- Develop new sampling algorithms based on the concept of the neighbourhood of influence.
Under-Sampling the Minority Class to Improve the Performance of Over-Sampling Algorithms in Imbalanced Data Sets

Romero F. A. B. de Morais (rfabm@cin.ufpe.br)
Germano C. Vasconcelos (gcv@cin.ufpe.br)

Center for Informatics
Federal University of Pernambuco

IJCAI Workshop on Learning in the Presence of Class Imbalance and Concept Drift, August 2017