Categorizing the feature space for two-class imbalance learning

Rosa Sicilia, Ermanno Cordelli, and Paolo Soda
Unit of Computer Systems and Bioinformatics
Department of Engineering
Università Campus Bio-Medico di Roma

The Problem
Class imbalance, a.k.a. class skew, refers to the case where certain prior probabilities of some classes are significantly lower than those of other classes.

Traditional machine learning algorithms are internally biased towards the majority class, producing poor predictive accuracy on the minority class.

Imbalanced learning Scenario

Class imbalance significantly affects the performance of classification algorithms. Imbalanced learning is the process of learning from data where the number of instances in the different classes is not equal.

Internal Approaches
The algorithm is tailored to imbalanced data exploiting specific knowledge of both classifier and application domain.

Cost-sensitive Learning
They consider the cost of wrong decisions and utilize a learner objective function sensitive to (class) costs.

Cost-sensitive Learning

A new technique to construct an ensemble of classifiers able to deal with binary imbalance learning tasks.

Contributions

A novel approach to characterize the feature space to detect reliable and unreliable configurations.

Materials and Methods

Building the Training sets
1. Each sample assigned to one of the 8 cases regardless of the label.
2. Compute the imbalance ratio among the samples in the j-th RSC class.
3. The number of classifiers is set according to the maximum $r_i$.
4. Each training set is composed of $\frac{|N_j|}{N}$ instances sampled with replacement from $N$ for each of the 8 RSC classes.
5. The final label is assigned by Majority Voting.

Experimental results

The proposed approach outperforms the competitors with a statistical significance difference on both metrics in most of the cases.

80%
Cases for Gmean

60%
Cases for IBA

Simple Bagging and Boosting can be more effective than using a specific method for class imbalance.

The proposed method beats the whole category of MES competitors.

Next Steps...

New way to construct and ensemble of classifiers for learning under class skew.

A novel method to categorize the feature space distinguishing reliable and unreliable configurations.

Promising performance: the proposal outperforms 15 competitors tested on 25 datasets.

Explore soft level combination strategies, rather than hard level ones.

Analyze different sample extraction procedures, rather than an exhaustive approach.

Investigate method performance with very large datasets.

Statistically assess relative degradation and recoveries among different methods.