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

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Imbalanced learning Scenario

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.

Internal Approaches

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

Data level Approaches

They modify the data distribution to create balanced datasets.

Cost-sensitive Learning

They consider the cost of wrong decisions and utilize a learner objective functions sensitive to (class) costs.

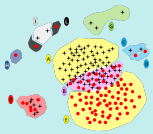
Ensemble Learning

They combine different balanced classifiers to get the final decision on each test sample.

Contributions

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

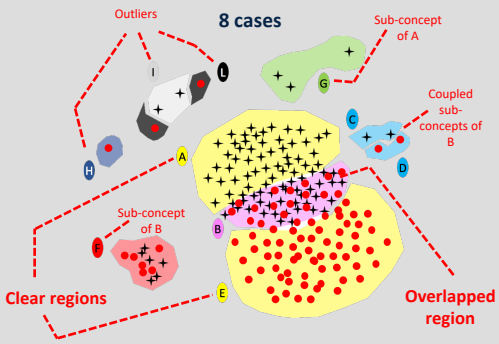
A novel approach to characterize the feature space to detect reliable and unreliable configurations.



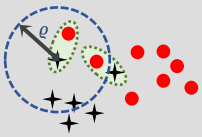
New algorithm to construct the training set of each base classifier so that it includes a proportion of positive and negative instances representing the different situations that can give rise to reliable or unreliable classifications.

Materials and Methods

Rule-based Space Chracterization



Meta-features



- Fraction of Tomek (fr_T)
- Number of same class samples (n_C)
- Fraction of same class samples (fr_C)

Building the Training sets

1. RSC — Each sample assigned to one of the 8 cases regardless of the label
2. $r^j = \frac{|N^j|}{|P^j|}$ — We compute the imbalance ratio among the samples in the j-th RSC class
3. C_1 C_2 ... C_{n_C} — The number of classifiers is set according to the maximum r^j
4. C_i training — Each training set is composed of $|N^j|/n_C$ instances sampled with replacement from N^j and P^j for each of the 8 RSC classes
5. The final label is assigned by Majority Voting

Experimental results

x25 Datasets

x15 Competitors

5-fold cross validation

C4.5 as base learner

Metrics

Gmean

$$g = \sqrt{acc^+ \cdot acc^-}$$

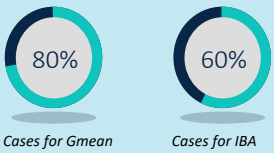
Accuracy is biased towards the majority class

Index of Balanced Accuracy

$$IBA = 1 + \alpha \cdot (acc^+ - acc^-) \cdot acc^+ \cdot acc^-$$

Iman-Davenport rank analysis	Method	g	IBA
	Proposal	12.58	12.54
Imbalanced Baselines	Imbalanced Classifier	4.82	4.3
	Bagging	7.8	6.8
	AdaBoost	9.1	8.02
Cost Sensitive	AdaBoostNC	9.06	7.5
	AdaC2-I	11.22	11.46
Boosting-based	EUSBoost	8.54	11.14
	MSMOTEBoost	11.04	10.88
	MSMOTEBagging	9.24	9.16
Bagging-based	OverBagging	11.96	10.52
	UnderBagging	9.46	11.78
	IIVotes	8.22	7.26
Ensemble	EasyEnsemble	8.32	9.48
	BalanceCascade	9.36	10.4
MES	MES-random	3.86	3.38
	MES-kmeans	1.42	1.38

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



Cases for Gmean *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.

TO DO

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

Analyse 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.