# Rare Class Prediction

## Research and ideas

### Rare Class Mining: Progress and Prospect a systematic Review

<https://ieeexplore.ieee.org/abstract/document/5344137>

Rare Class vs. Imbalanced Datasets

Imbalanced Dataset 🡪 focusses on imbalanced distribution of classes

Rare Class mining 🡪 rare class mining puts more emphasis on the detection of the rare class

very limited work has been done on feature selection techniques in the rare class mining domain

Good Practices

### Sampling-Based Methods

General:

* Ratio between 2:1 or 3:1 have best performance
* Typically sampling techniques are domain dependent
* These methods are apparently not applied in time series

Undersampling – discard majority class

* Results in information loss and bias

Oversampling – duplicate minority class

* Increases computational complexity.
* May causes overfitting.

Heuristic Sampling Methods

* discarded are those noisy, redundant or borderline ones

### Cost-Sensitive Learning

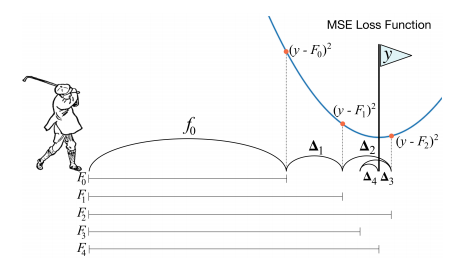
Misclassification of rare class is penalized.

It is difficult to set the right penalty as domain expertise is required

Sun et al use the genetic algorithm to find the optimal misclassification cost

### Algorithms for Rare Class Mining

#### Boosting Algorithms



There is adapted versions of boosting for better performance

#### Rule-Based Algorithms

two-phase rule-induction approach

* first phase, rules that have high support and reasonable accuracy are discovered
* second phase, rules able to remove the false positive samples are developed

#### Division-Based Algorithms

* Clustering of rare class

Research Topics derived from summary

how to select the best sampling technique and parameter combination is still an important research topic

### Anomaly Detection

Typically dealing with time series

Three broad categories of anomaly detection techniques exist.[[4]](https://en.wikipedia.org/wiki/Anomaly_detection#cite_note-ChandolaSurvey-4) **Unsupervised anomaly detection** techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. **Supervised anomaly detection** techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification) problems is the inherent unbalanced nature of outlier detection). **Semi-supervised anomaly detection** techniques construct a model representing normal behavior from a given *normal* training data set, and then test the likelihood of a test instance to be generated by the learnt model.

### Time Series Anomaly Detection Using Convolutional Neural Networks and Transfer Learning

<https://arxiv.org/pdf/1905.13628.pdf>

Anomaly detection using CovNets (U-Nets and MU-Nets)

* Potential research areas are comparison Cov-Nets and RNN
* Or Extensive work on how to pretrain

Comparison of methods

https://arxiv.org/pdf/2004.00433v1.pdf

Potential Datasets

https://github.com/robertoamansur/rare\_event\_pred\_maintanance