ON THE CLASS IMBALANCE PROBLEM

XINJIAN GUO,YILONG YIN1 , CAILING DONG, GONGPING YANG, GUANGTONG ZHOU

FOURTH INTERNATIONAL CONFERENCE ON NATURAL COMPUTATION

MOTIVATION

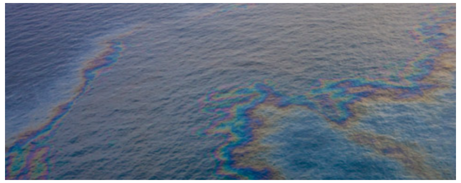
- I chose to review this research as I am working on a Kaggle problem related to fraud detection
  - I want to boost my performance via some modern methods
- Fraud detection is one of the main practical examples of a class imbalance problem
- This paper reviews some methods for tackling class imbalance and has around 200 citations

ABSTRACT

- Class Imbalance is a hot topic in Machine Learning
- What is the Class Imbalance Problem?
  - Nearly all of the target labels are one class while extremely few belong to the other class
  - Example: 99.9987 % Cases are normal - .0013 Cases are Fraud
- Standard Machine Learning Algorithms can become overwhelmed by the majority class negatively impacting their ability to generalize
- This paper reviews recent academic activities related to class imbalance, various methods for handling the problem, and some future directions (2008)

INTRODUCTION

- Many machine learning algorithms assume that target classes share prior probabilities
- Many real world applications are however imbalanced (Nearly all examples belonging to one class)
  - Oil-spill detection
  - Network intrusion detection
  - Fraud detection
  - Medical Diagnosis
- There are two general cases for Class Imbalance:
  - Data is naturally unbalanced (Credit Fraud or Rare Disease)
  - Data is expensive to obtain for the minority class (Shuttle Failure)
- This section also reviews the structure of the research paper



ACADEMIC ACTIVITIES ON THE CLASS IMBALANCE PROBLEM

- **Academic activities related to the Class Imbalance Problem cited in the research paper:**
- First workshop dedicated to class imbalance- American Association for Artificial Intelligence conference (2000)
  - Important issues like application domains dealing with imbalanced data sets, evaluation measures, discussions over re-sampling methods, cost-sensitive learning, among other topics were discussed
- The second workshop dedicated to class imbalance- International Conference on Machine Learning (2003)
  - Most research was guided by the first conference but ROC or Cost curves were used as evaluation metrics instead of accuracy, re-sampling was presented and also debated in many papers
- The Sixth issue of SIGKDD Exploration was dedicated entirely to the class imbalance problem
  - Papers in the volume discussed research on learning from imbalanced datasets, issues of sampling, feature selection, and one-class learning

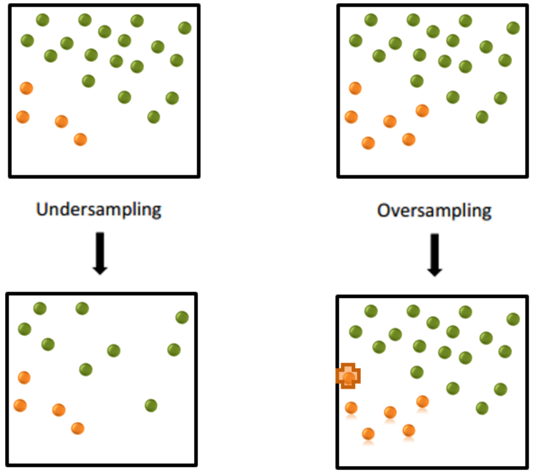
REMEDIES FOR THE CLASS IMBALANCE PROBLEM

- There are four main methods for dealing with class imbalance depending on what phase of learning:
  - Changing Class Distributions
  - Feature Selection
  - Modifications within the Classifier itself
  - Ensemble Learning Methods

CHANGING CLASS DISTRIBUTIONS

3.1 CHANGING CLASS DISTRIBUTIONS

- **Changing Class distributions – Changing the nature of the data distribution by removing or creation of synthetic examples (Data Level)**
- Methods for creating balanced distributions include :
  - Under-Sampling the majority class
  - Over-sampling the minority class
  - Combination of both of these methods / more advanced methods

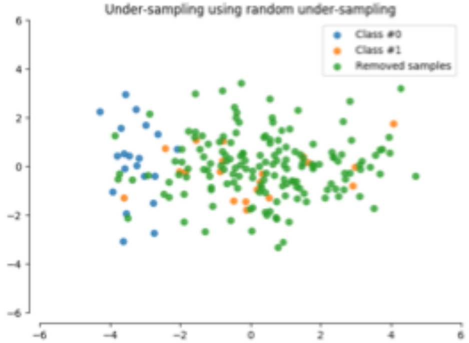


- **Conclusion from this section :**
  - "Various strategies for learning from imbalanced data sets were compared, and it concluded that under-sampling and over-sampling are very effective methods for dealing with the class imbalance problem."

3.1.1 UNDER-SAMPLING

- **Random under-sampling** – The most naïve under-sampling method, tries to balance class distributions through random elimination of majority class samples
  - Could potentially lead to discarding of useful data

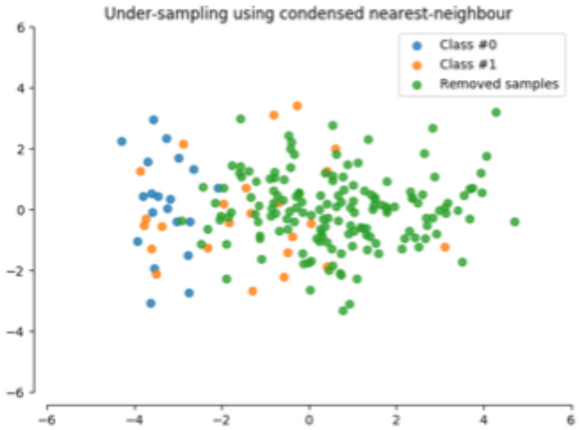
- Two Noise Model Hypothesis related to Under-sampling include:
  - Hypothesis 1- (Noise) - examples near the decision boundary between two classes
  - Hypothesis 2- (Noise) - examples with more neighbors of differing labels



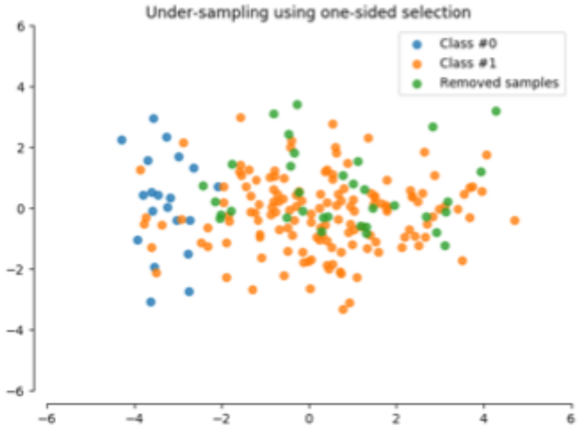
Source: scikitlearn

3.1.1 UNDER-SAMPLING

- **Decision Boundary Related**
- **Condensed Nearest Neighbor Rule (CNN)** – Nearest Neighbor based sampling method, goal is to reduce data and maintain samples close or near to the decision boundary based on Bayesian risk. CNN is effective when binary classes do not overlap much.
- **OSS** –The goal is to eliminate all the examples from the majority class that are distant from the decision boundary as these may be less relevant for learning
  - Randomly selects one example from majority class and all examples from the minority class and puts these in E'
  - Then uses 1-NN to classify the examples in E (Majority Class Examples)
  - Every misclassified example is moved from E to E'



Under-sampling using condensed nearest-neighbour

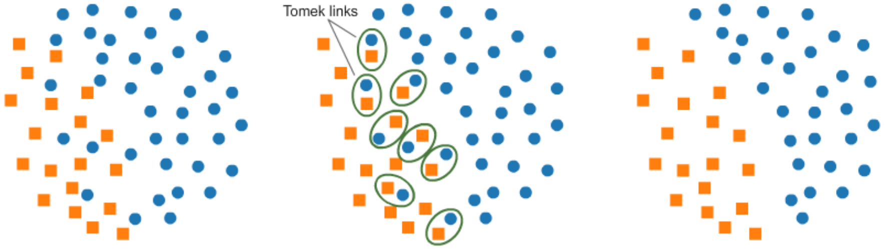


Under-sampling using one-sided selection

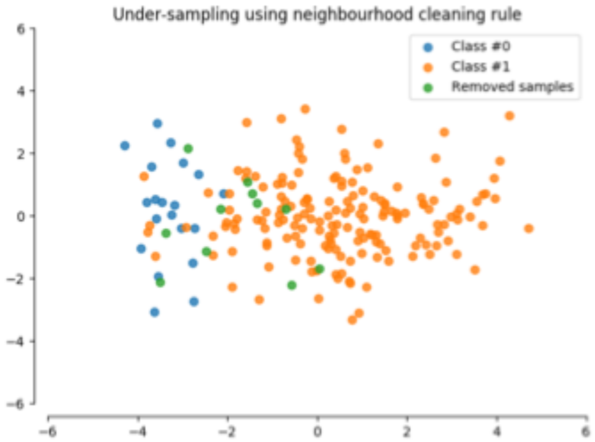
Source: scikitlearn

3.1.1 UNDER-SAMPLING

- **Neighbors of Differing Labels Approaches**
- **Wilson's Edited Nearest Neighbor Rule (ENN)** – removes any example whose class labels differ from the class of at least two of its three neighbors.
- **Neighborhood Cleaning Rule (NCL) –** Deals with Majority and Minority sets differently when under-sampling. NCL Uses ENN to clean the examples.
  - "If Ei belongs to the majority class and the classification given by its three nearest neighbors contradicts the original class of Ei, then Ei is removed. If Ei belongs to the minority class and its three nearest neighbors misclassify Ei, then the nearest neighbors that belong to the majority class are removed. "
- **Tomek Links –** examples near the borderline are considered more relevant –
  - "Given two examples Ei and Ej belonging to different classes, and d(Ei, Ej) is the distance between Ei and Ej; a (Ei, Ej) pair is called a Tomek link if there is not an example E1, such that d(Ei, E1) < d(Ei, Ej) or d(Ej , E1) < d(Ei, Ej).
  - If two examples form a Tomek Link these observations can be considered noise and removed
- *Tomek Link, ENN, and NCL are considered highly time consuming methods as they require calculation of nearest neighbors



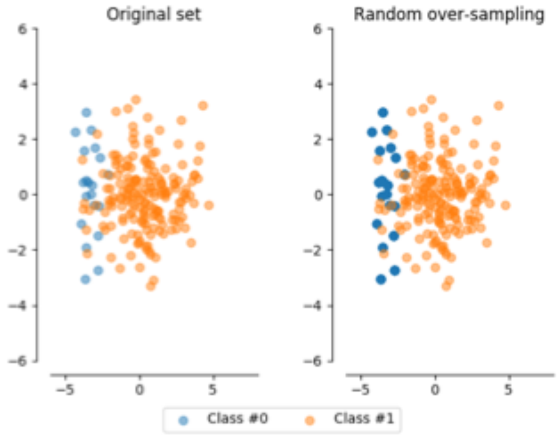
Source: Kaggle

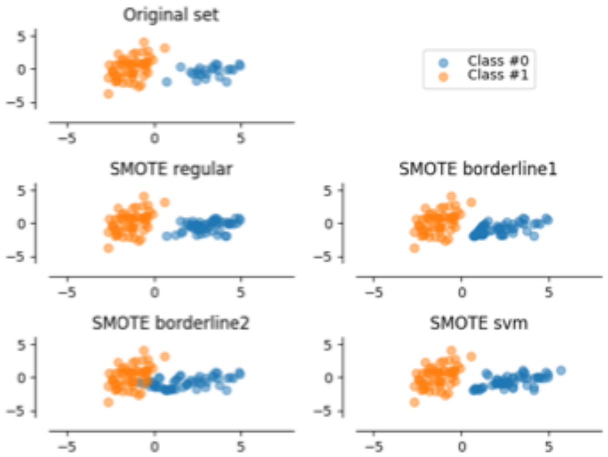


Source: scikitlearn

3.1.2 OVER-SAMPLING

- **Random over-sampling –** Similar to Random Under-Sampling but this time aims to balance class distributions through random replications of the minority class
  - Can lead to overfitting since it makes exact copies of minority class
  - Makes learning process more time consuming
- There are several over-sampling methods based on **SMOTE (Synthetic Minority Oversampling Technique)**
  - SMOTE generates synthetic examples to over-sample the minority class, by generating instead of replicating data to avoid the overfitting problem.
- **Borderline_SMOTE –** This was proposed to address the problem of examples near the border line being more easily misclassified. This method only oversamples the border examples from the minority class or a subset of the minority class.
  - According to literature, this has had increased performance in F-value and TP rate in comparison to Random over-sampling and generic SMOTE

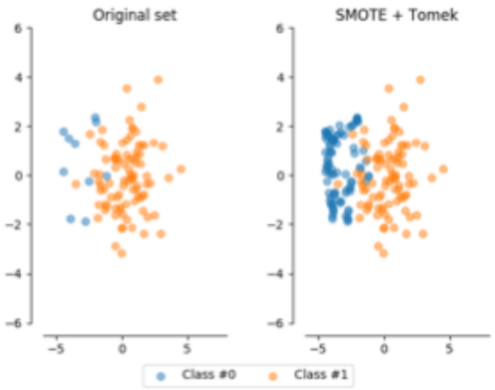


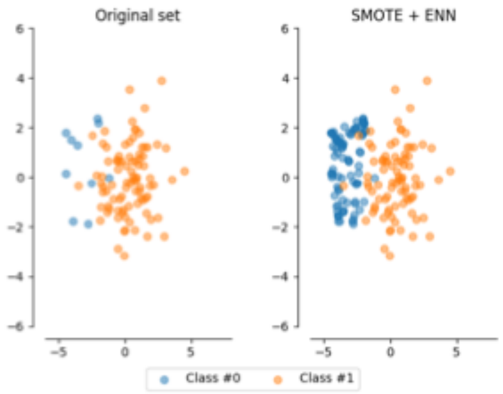


Source: scikitlearn

3.1.3 ADVANCED SAMPLING

- **Advanced Sampling Methods differ from under-sampling and over-sampling methods as they do re-sampling based on the results of preliminary classifications**
- **Boosting –** iterative algorithm that places weights on the training distributions each iteration. After each iteration boosting increases the weights associated with incorrectly classified examples and decreased the weights associated with correctly classified examples.
- **Preliminary Classification (OSPC)** - Another method proposed by Han et. al. was based on preliminary classification of the test data, test data that were predicted to belong to minority class were reclassified to improve performance of the minority class.
- **Another Key Point -** "When the datasets are severely skewed, under-sampling and over-sampling methods are often combined to improve generalization of the learner"
  - Ex 1: SMOTE combined with Tomek Link
  - Ex 2: SMOTE combined with ENN





FEATURE SELECTION

3.2 FEATURE SELECTION

- **Feature Selection is also known as variable selection- The selection of the most suitable features for a model (Feature Selection Level)**
- Most work in feature selection cited in this paper has been related to the text classification and Web Categorization domains
- A few papers referenced here:
  - Zhang et. Al. proposed a feature selection framework that selects features separately for positive and negative classes
  - Putten and Someren analyzed the COIL 2000 data sets using the bias-variance decomposition –
    - They concluded that feature selection in such domains is even more important than the learning method selected

CLASSIFIERS LEVEL

CLASSIFIERS LEVEL

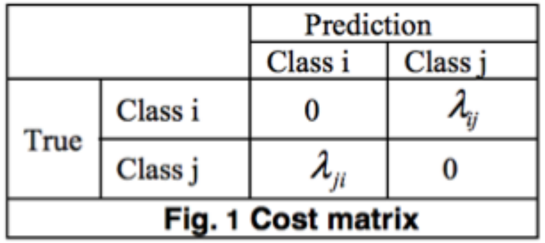
- Manipulating classifiers internally
- Cost-Sensitive Learning
- One-class learning

3.3.1 MANIPULATING CLASSIFIERS INTERNALLY

- **Below are methods for modifying classifiers to improve their performance on imbalanced datasets (Classifier Level)**
  - **C4.5-** over-sampling did not yield much performance improvements with C4.5, adjustment of parameter settings are required to increase the influence of tree pruning and overfitting avoidance
  - **Naïve Bayes / Neural Networks** – some classifiers provide a score that represents how much an example belongs to a certain class. This ranking can be used to produce several classifiers based on thresholds.
  - **kNN** – A weighted distance function was proposed to modify kNN's behavior on imbalanced datasets. Weights based on which class an example belongs to.
  - **SVM –** Methods for dealing with class imbalance in SVM include strategies to move the hyperplane closer to the minority-class, this can be done through changing the kernel function or different penalty constants for different classes of data

3.3.2 COST-SENSITIVE LEARNING

- **Cost-Sensitive Learning methods are based the concept of cost where certain errors have different costs associated with them**
  - Adding costs to the decision making process is another way to improve classifiers performance
- **Cost Model:**
  - The cost matrix is usually expressed in terms of average misclassification cost for the problem
  - The diagonal (correct) terms are set to 0 (Correct classification has no cost)
  - The goal is to minimize the cost (Lowest misclassifications)



		Prediction	
		Class i	Class j
True	Class i	0	λ_{ij}
	Class j	λ_{ji}	0

Fig. 1 Cost matrix

- **MetaCost –** Another method to make a classifier cost-sensitive
- **AdaCost –** (Modification to make AdaBoost's weight-update rule Cost Sensitive) – weights assigned to to the rare class that are misclassified have higher cost then those belonging to the common class – Has proven to outperform AdaBoost
  - AdaBoost is a form of boosting

3.3.3 ONE-CLASS LEARNING

- **One-class learning are recognition based approaches where the model is generated based on the target class alone based on similarity values**
  - (Goal is to learn only the minority class)
- Brute, Shrink, and Ripper are three methods used in many modern applications that still train using examples belonging to all classes
  - Brute was utilized to look for flaws in Boeing Manufacturing Process
  - Shrink was used to detect rare oil spills from satellite radar images
- One-Class Learning methods have been shown to be competitive especially in domains where the datasets are highly dimensional and extremely unbalanced

ENSEMBLE LEARNING METHODS

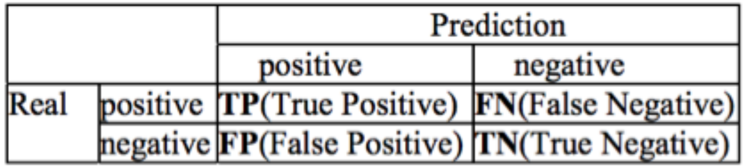
3.4 ENSEMBLE LEARNING METHODS

- **Ensemble Learning Methods combine the results of many classifiers to handle the class imbalance problem.**
  - Combination of Multiple Models to generate a final classification
  - Perhaps the most known approaches are bagging and boosting
- There are many approaches covered in the literature including:
  - **AdaBoost**
  - **Rare-Boost**
  - **SMOTEBoost**
  - **MetaCost**
  - **Stacking-Bagging**
  - **Combination of C4.5 and kNN**

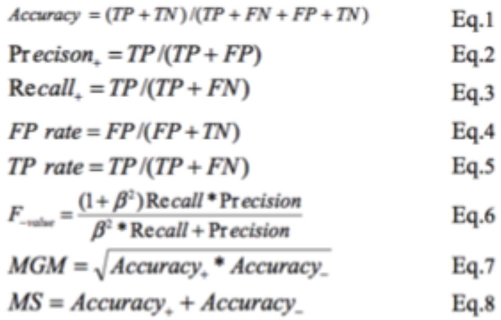
EVALUATION METRICS

- Accuracy is usually the most common evaluation metric in traditional classification problems but is not suitable to evaluate imbalanced datasets
  - Example: If you build a classifier that classifies all examples to the majority class then you will have around 99 % Accuracy most of the time. This good accuracy score does not take into account the weight of the minority class.
- Several Metrics are popular in the domain of imbalanced classification – Fall into 2 categories
  - Metrics Based on Confusion Matrix
    - ~~Accuracy~~
    - Precision/Recall
    - FP Rate
    - TP Rate
    - ROC and AUC
  - Metrics based on Accuracy or Precision/Recall
    - F-Values
    - Maximum Geometry Mean (MGM)
    - Maximum Sum (MS)

Table 1. Confusion matrix



		Prediction	
		positive	negative
Real	positive	**TP**(True Positive)	**FN**(False Negative)
	negative	**FP**(False Positive)	**TN**(True Negative)



$$Accuracy = (TP + TN)/(TP + FN + FP + TN) \qquad \text{Eq.1}$$

$$\mathrm{Pr}ecison_+ = TP/(TP + FP) \qquad \text{Eq.2}$$

$$\mathrm{Re}call_+ = TP/(TP + FN) \qquad \text{Eq.3}$$

$$FP\ rate = FP/(FP + TN) \qquad \text{Eq.4}$$

$$TP\ rate = TP/(TP + FN) \qquad \text{Eq.5}$$

$$F_{-value} = \frac{(1+\beta^2)\mathrm{Re}call * \mathrm{Pr}ecision}{\beta^2 * \mathrm{Re}call + \mathrm{Pr}ecision} \qquad \text{Eq.6}$$

$$MGM = \sqrt{Accuracy_+ * Accuracy_-} \qquad \text{Eq.7}$$

$$MS = Accuracy_+ + Accuracy_- \qquad \text{Eq.8}$$

Fig.2. Evaluation metrics based on confusion matrix

RELATIONS TO OTHER PROBLEMS

- Some datasets are able to produce good classifiers despite having highly imbalanced training sets
  - The distributions within each class of the data are also considered important
  - Prati et. al. [54] - Problem is not solely related to class imbalance but degree of data overlapping among the classes
- Data duplication is generally harmful
- Misclassification usually occurs near class boundaries
- Relation between class imbalance and training set size - Increasing the size of the training set always leads to improved classifier importance.

CONCLUSION

- This paper reviewed many literatures related to the handling of imbalanced classification
- This research will be useful in my future applications of Machine Learning for Imbalanced Class Problems