Caveats and Techniques to Deal with Imbalanced Datasets

ML is now everywhere

- Increase of storage capacity → easy to build large datasets
e.g. companies can store activities of clients with a variety of attributes

- Widely available code for ML techniques → Easy to use by laymen
  ➢ ML used widely in a variety of domains

- One hopes that ML will solve problems with which companies struggle
  (vast amount of data, very noisy, high-dimensional)
ML techniques assume balanced datasets

ML algorithms assume that data is balanced
➢ In classification: comparative number of instances of each class

What would be the result of running SVM on the imbalanced dataset?
Imbalance is Everywhere

- Between-class
- Within-class
- Intrinsic and extrinsic
- Relativity and rarity
- Imbalance and small sample size
Imbalanced Data: Example

Finance:
Number of clients who closed account / nm clients who did not: 1%

Robotics:
Number of points where the two arms intersect / nm points with no intersection: 4-5%

One has usually much less datapoints from the adverse class. This is unfortunate as we care a lot about avoiding misclassifying elements of this class.
Imbalanced Data: Example

Intrinsic and extrinsic imbalance

Intrinsic:
- Imbalance due to the nature of the dataspace

Extrinsic:
- Imbalance due to time, storage, and other factors
- Example:

Data transmission over a specific interval of time with interruption

What would be the effect on SVR?
Imbalanced Data: Example

Poor interpolation with missing data
Types of Imbalance

Balanced between-class datasets but white dataset not representative with more data is some regions
Imbalance ≠ rarity

*Relative imbalance and absolute rarity*

\[ Q: 1,000,000 : 1,000 = 1,000 : 1 \]

- The minority class may be outnumbered, but not necessarily rare
- Therefore they can be accurately learned with little disturbance
Imbalance and curse of dimensionality

*Imbalanced data with small sample size*

- Data with high dimensionality and small sample size
  - Face recognition, gene expression

- Challenges with small sample size:
  1. Embedded absolute rarity and within-class imbalances
  2. Failure of generalizing inductive rules by learning algorithms
     - Difficulty in forming good classification decision boundary over *more* features but *less* samples
     - Risk of overfitting
Approaches to learning with imbalanced datasets
Learning with imbalanced datasets

Two main approaches

Act on the data
Sampling Methods

Act on the cost function
Cost-Sensitive Methods
Sampling methods

1. If data is imbalanced...
2. Modify data distribution
3. Create balanced dataset

Create balance though sampling
Sampling Methods

Compensate the lack of data by:

- **Increase Dataset**
  - Generate new data points for the smallest class

- **Decrease Dataset**
  - Remove redundant datapoints from the largest class
Undersampling

Remove redundant datapoints

Looses statistics – good only if enough datapoints on undersampled class and for low dimensional datasets
Oversampling

Pick neighbour and create new datapoint

Risk overfitting, especially if one does this for points that are noise
SMOTE: Synthetic minority oversampling technique

Generate new samples inbetween existing datapoints based on their local density and their borders with the other class. Can use cleaning techniques (undersampling) to remove redundancy in the end.

- No Neighbors of the same class → **noise**
- Several Neighbors of the same class
- Surrounded by the other class → **in danger**
- Surrounded only on one side by the other class → **safe**
Cost-Sensitive Methods

Instead of modifying data...

Considering the cost of misclassification

Utilize cost-sensitive methods for imbalanced learning
Cost-Sensitive Learning Framework

- Define the cost of misclassifying a majority to a minority as $C(Min, Maj)$
- Typically $C(Maj, Min) > C(Min, Maj)$
- Minimize the overall cost - usually the Bayes conditional risk - on the training data set

$$R(i|x) = \sum_j P(j|x)C(i,j)$$

Fig. 7. Multiclass cost matrix.
Cost-Sensitive Dataspaces Weighting with Adaptive Boosting

• Iteratively update the distribution function $D_t$ of the training data according to the error of the current hypothesis $h_t$ and cost factor $C_i$

  • Weight updating parameter $\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right)$

  • Error of hypothesis $h_t$: $\varepsilon_t = \sum_{i: h_t(x_i) \neq y_i} D_t(i)$
How to treat Imbalanced Datasets with SVM

- SVM with asymmetric misclassification cost
  
  \[ \text{minimize} \quad \frac{1}{2} \|w\|^2 + \frac{C}{M} \sum_{i=1}^{M} \beta_i \left( \xi_i + \xi_i^* \right) \]

- SVM class boundary adjustment
  
  \[ y = \text{sgn} \left( \sum_{i=1}^{M} \alpha_i k(x^i, x) + b \right) \]
How to treat Imbalanced Datasets with SVM

Active Learning Methods

• SVM-based active learning

![Data imbalance ratio within and outside the margin][1]

Fig. 8. Data imbalance ratio within and outside the margin [98].

• Active learning with sampling techniques
  • Undersampling and oversampling with active learning for the word sense disambiguation (WSD) imbalanced learning
  • New stopping mechanisms based on maximum confidence and minimal error
  • Simple active learning heuristic (SALH) approach
Taking into account imbalanced datasets in the assessment of performance

Traditional accuracy measure are sensitive to the data distribution

<table>
<thead>
<tr>
<th>Hypothesis output</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
</tr>
<tr>
<td>Y</td>
<td>TP (True Positives)</td>
</tr>
<tr>
<td>N</td>
<td>FN (False Negatives)</td>
</tr>
</tbody>
</table>

Column counts:  

- \( P_C \)
- \( N_C \)

\[
\text{Accuracy} = \frac{TP + TN}{P_C + N_C}
\]

\[
\text{Error Rate} = 1 - \text{accuracy}
\]

The F-measure is better adapted as it evaluated performance on one class.

\[
F\text{-Measure} = \frac{(1 + \beta)^2 \cdot \text{Recall} \cdot \text{Precision}}{\beta^2 \cdot \text{Recall} + \text{Precision}}
\]

\[
\beta = 1, \text{ usually}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]
Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning

Guillaume Lemaître
Pariteal team, Inria, CEA, Université Paris-Saclay
1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France

Fernando Nogueira
ShoppeAI
488 Wellington Street West, Suite 304, Toronto, Ontario M5V 1E3, Canada

Christos K. Aridas
Computational Intelligence Laboratory
Department of Mathematics
University of Patras
GR-26504 Patras, Greece

Editor: Geoff Holmes