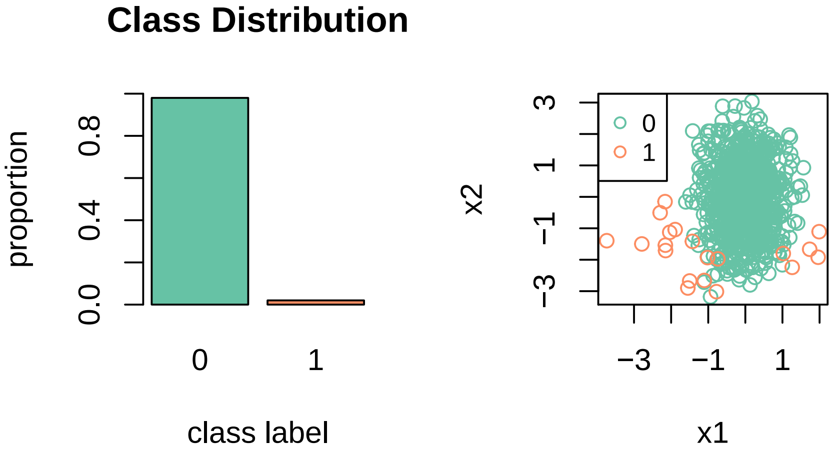
Overview of Imbalanced Classification

Diane Lu

Classifying under the setting that the number of samples in different classes are very different.



Many Real-World Situations

Imbalanced data is very common in various real-world situations. And the problem becomes more challenging (overfitting) in the presence of *rare events*, below are some of the examples:

- Cancer diagnosis
- Spam detection
- Fraud credit card transaction detection
- Natural disasters prediction

What Could Go Wrong With Usual Metric?

- If we're using "accuracy" as the performance measure...
- Dummy classifier that classifies everything to the majority could still maintain high accuracy.
- Not useful?

Confusion Matrix

It seems that "accuracy" isn't giving us the whole picture, since our primary goal is to correctly identify the minority class.

A simple but effective evaluation criterion for skewed class distribution is the confusion matrix.

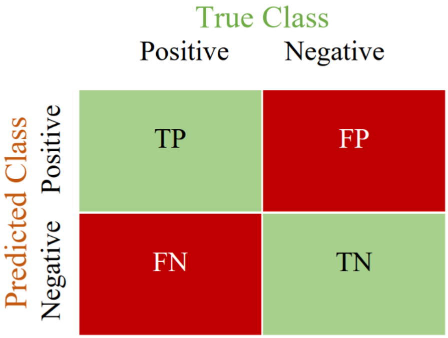


Figure 1: Confusion Matrix

Accuracy, Precision, Sensitivity, Specificity

Accuracy: (TP+TN)/(TP+FP+TN+FN), how many samples are correctly predicted out of all samples.

Precision: TP/(TP+FP), how many samples are truly positive out of all positive predictions.

Sensitivity (Recall): TP/(TP+FN), how many samples are predicted positive out of all true positive samples. A.k.a True Positive Rate.

Specificity: TN/(TN+FP), how many samples are predicted negative out of all true negative samples. A.k.a. True Negative Rate.

For the ideal case, where FN and FP are 0, the above 4 metrics would be 1.

ROC Curve

- receiver operating characteristic curve
- illustrate the diagnostic ability of a binary classifier as its discrimination threshold is varied
- access the tradeoff between sensitivity (TPR) and specificity (1-FPR)

Area Under the ROC Curve (AUC)

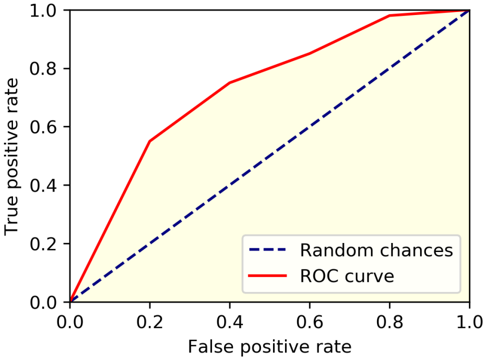


Figure 2: ROC curve

- Want a classifier with a large Area Under the ROC Curve (AUC).

Image source: Huy Bui

ROC curves for multi-class classification?

One vs all method.

If you have three classes named A, B and C, you can have one ROC for each of the following cases:

- A vs. B and C
- B vs. A and C
- C vs. A and B

Common Re-sampling Methods

- Undersampling: random undersample the majority class
- Oversampling : random oversample the minority class
  - SMOTE(Synthetic Minority Over-Sampling Technique): interpolate between nearby minority samples to avoid overfitting (Chawla et al. (2002), image source: He and Garcia (2009))

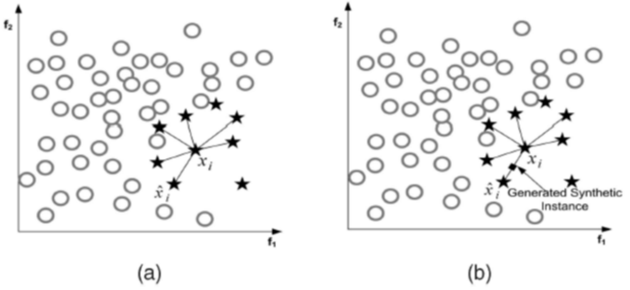


Fig. 3. (a) Example of the K-nearest neighbors for the x_i example under consideration ($K = 6$). (b) Data creation based on euclidian distance.

- Place more weights on the minority class (or larger penalty constant), forcing the classifier to correctly classify on the minority samples. Ex: Cost-Sensitive Support Vector Machines (SVM) (Veropoulos et al. (1999) and Wu and Chang (2003)), Label-Distribution-Aware Margin Loss (Cao et al. (2019))

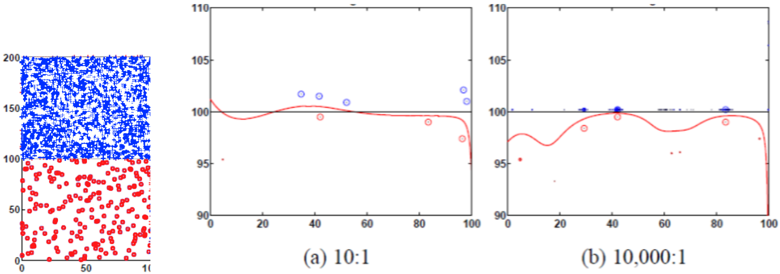


Figure 3: SVM with Different Imbalanced Ratio

Common Ensemble Methods

- MetaCost: bagging + Cost Matrix (Domingos (1999))
- SMOTEBagging: SMOTE + bagging (Wang and Yao (2009))
- UnderBagging: random undersampling + bagging (Barandela, Valdovinos, and Sánchez (2003))
- SMOTE-Boost: SMOTE + boosting (Chawla et al. (2003))
- RUSBoost: random undersampling + boosting (Seiffert et al. (2009))

Deep Learning Methods

- Focal Loss: similar to re-weighting scheme. It changes the loss function by giving less weights on the samples that are classified correctly (Lin et al. (2017))
- Data Generation through conditional generative adversarial networks (cGAN): similar to oversampling, but uses cGAN to generate more minority samples. (Douzas and Bacao (2018))

References I

Barandela, Ricardo, Rosa Maria Valdovinos, and José Salvador Sánchez. 2003. "New Applications of Ensembles of Classifiers." *Pattern Analysis & Applications* 6 (3): 245–56.

Cao, Kaidi, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. "Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss." In *Advances in Neural Information Processing Systems*, 1567–78.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57.

Chawla, Nitesh V, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. 2003. "SMOTEBoost: Improving Prediction of the Minority Class in Boosting." In *European Conference on Principles of Data Mining and Knowledge Discovery*, 107–19. Springer.

References II

Domingos, Pedro. 1999. "Metacost: A General Method for Making Classifiers Cost-Sensitive." In *Proceedings of the Fifth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 155–64.

Douzas, Georgios, and Fernando Bacao. 2018. "Effective Data Generation for Imbalanced Learning Using Conditional Generative Adversarial Networks." *Expert Systems with Applications* 91: 464–71.

He, Haibo, and Edwardo A Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–84.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. "Focal Loss for Dense Object Detection." In *Proceedings of the Ieee International Conference on Computer Vision*, 2980–8.

References III

Seiffert, Chris, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2009. "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40 (1): 185–97.

Veropoulos, Konstantinos, Colin Campbell, Nello Cristianini, and others. 1999. "Controlling the Sensitivity of Support Vector Machines." In *Proceedings of the International Joint Conference on Ai*. Vol. 55.

Wang, Shuo, and Xin Yao. 2009. "Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models." In *2009 Ieee Symposium on Computational Intelligence and Data Mining*, 324–31. IEEE.

Wu, Gang, and Edward Y Chang. 2003. "Class-Boundary Alignment for Imbalanced Dataset Learning." In *ICML 2003 Workshop on Learning from Imbalanced Data Sets*.