Learning with class-imbalanced datasets

Nicolás García-Pedrajas

Computational Intelligence and Bioinformatics Research Group

April 20, 2020

Table of contents

Introduction

- High imbalance occurs in real-world domains where the decision system is aimed to detect a rare but important case.
- Exists in many real-world domains
  - ✓ Spotting unreliable telecommunication customers
  - ✓ Detection of oil spills in satellite radar images
  - ✓ Learning word pronunciations
  - ✓ Text classification
  - ✓ Detection of fraudulent telephone calls
  - ✓ Information retrieval
  - ✓ Filtering tasks

Introduction

- A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels.
  - ✓ At the data level
  - ✓ At the algorithmic level
  - ✓ Ensemble approaches (combining methods)

Data level methods for handling imbalance

- Data level solutions include many different forms of
  - ✓ Re-sampling such as random oversampling with replacement
  - ✓ Random undersampling
  - ✓ Directed oversampling
  - ✓ Directed undersampling
  - ✓ Oversampling with informed generation of new samples
  - ✓ Combinations of the above techniques

Undersampling

- Random under-sample
  - ✓ It is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples.
- Tomek Links[Tomek, 1976]
  - ✓ E_i and E_j belonging to different classes
  - ✓ $d(E_i, E_j)$ is the distance between E_i and E_j
  - ✓ (E_i, E_j) pair is called a Tomek link if there is not an example E_1, such that $d(E_i, E_1) < d(E_i, E_j)$ or $d(E_j, E_1) < d(E_i, E_j)$.
- Kubat and Matwin [1997]
  - ✓ Randomly draw one majority class example and all examples from the minority class and put these examples in E'.

Oversampling

- Random over-sampling
  - ✓ It is a non-heuristic method that aims to balance class distribution through the random replication of minority class examples
- Chawla et al. [2002]
  - ✓ Synthetic Minority Over-sampling Technique (SMOTE)
  - ✓ SMOTE generates synthetic minority examples to over-sample the minority class

Feature selection for imbalanced datasets

- Zheng et al. [2004] suggest that existing measures used for feature selection are not very appropriate for imbalanced data sets
- Feature selection framework
  - ✓ Selects features for positive and negative classes separately and then explicitly combines them
- The authors show simple ways of converting existing measures so that they separately consider features for negative and positive classes

Algorithm level methods for handling imbalance

- Drummond and Holte [2003] report that when using C4.5's
  - ✓ Oversampling is surprisingly ineffective, often producing little or no change in performance in response to modifications of misclassification costs and class distribution.
- Barandela et al. [2003] used in the classification phase of k-NN
  - ✓ The basic idea behind this weighted distance is to compensate for the imbalance in the training sample without actually altering the class distribution

Algorithm level methods for handling imbalance
SVM

Another approach to dealing with imbalanced datasets using SVM biases the algorithm so that the learned hyperplane is further away from the positive class

Threshold method

- Some classifiers, such as the Naïve Bayes classifier or some Neural Networks, yield a score that represents the degree to which an example is a member of a class
- The threshold can be adjusted to deal with class-imbalance
- Such ranking can be used to produce several classifiers, by varying the threshold of an example pertaining to a class.

One-class learning

- Raskutti and Kowalczyk [2004]
  - ✓ Useful when used on extremely unbalanced data sets composed of a high dimensional noisy feature space.
- An interesting aspect of one-class (recognition-based) learning is that, under certain conditions such as multi-modality of the domain space
- One class approaches to solving the classification problem may in fact be superior to discriminative (two-class) approaches (such as decision trees or Neural Networks)

Cost sensitive learning

Cost model takes the form of a cost matrix, where the cost of classifying a sample from a true class j to class i corresponds to the matrix entry λ_{ij}

Combining methods

- A mixture-of-experts approach has been used to combine the results of many classifiers, each induced after over-sampling or under-sampling the data with different over/under-sampling rates.
- Another method that uses this general approach employs a progressive-sampling algorithm to build larger and larger training sets

Combining methods

- Domingos [1999]: MetaCost method for making a classifier cost-sensitive.
- Joshi et al. [2001]: Rare-Boost scales false-positive examples in proportion to how well they are distinguished from true-positive examples and scales false-positive examples in proportion to how well they are distinguished from true-negative examples
- Chawla et al. [2003]: SMOTEBoost adapt SMOTE method to build ensembles for class-imbalance datasets

Evaluation metrics

- TP and TN denote the number of positive and negative examples that are classified correctly
- FN and FP denote the number of misclassified positive and negative examples respectively
  - ✓ Accuracy = (TP+TN)/(TP+FN+FP+TN)
  - ✓ FP rate = FP/(TN+FP)
  - ✓ TP rate = Recall = TP/(TP+FN)
  - ✓ Precision = TP/(TP+FP)
  - ✓ F-value= (1+β^2)Recall * Precision / β^2 Recall + Precision
    - Usually $\beta = 1$

Evaluate the performance of classifiers in learning

- Minimum Cost criterion (MC)
- Maximum Geometry Mean (MGM)
- Maximum Sum (MS)
- Receiver Operating Characteristic (ROC) analysis.

Minimum cost criterion (MC)

- Bradley [1997]
  - ✓ The MC criterion minimizes the cost measured by $Cost = FP \times CFP + FN \times CFP$
  - ✓ CFP is the cost of a false positive
  - ✓ CFN is the cost of a false negative
- However,the cost of misclassification is generally unknown in real cases, this restricts the usage of this measure

Maximum Geometry Mean (MGM)

- Kubat and Matwin [1997]
  - ✓ Accuracy on the majority class and the minority class
  - ✓ The criterion of MGM maximizes the geometric mean of the accuracy, but it contains a nonlinear form, which is not easy to be automatically optimized

Maximum sum (MS)

- Grzymala-Busse et al. [2003]
  - ✓ Accuracy on the majority class and the minority class
  - ✓ MS maximizing the sum of the accuracy on the positive class and the negative class (or maximizing the difference between the true-positive and false-positive probability) , is a linear form

Receiver Operating Characteristic (ROC)

- Bradley [1997]
- Perhaps the most common metric is ROC analysis and the associated use of the area under the ROC curve (AUC) to assess overall classification performance

Other problems related with imbalance

- Prati et al. [2011]
  - ✓ Developed a systematic study aiming to question whether class imbalances hinder classifier induction or whether these deficiencies might be explained in other ways.
- Their study was developed on a series of artificial data sets in order to fully control all the variables they wanted to analyze

Other problems related with imbalance

- A number of papers discussed interaction between the class imbalance and other issues
  - ✓ Japkowicz and S.Shaju [2002]: Small disjunct
  - ✓ Rare cases problems
  - ✓ Data duplication
  - ✓ Visa and Ralescu [2005]: Overlapping classes
- It was also found that data duplication is generally harmful, although for classifiers such as Naïve Bayes and Perceptrons with Margins, high degrees of duplication are necessary to harm classification

Other problems related with imbalance

- Jo and Japkowicz [2004]
  - ✓ experiments suggest that the problem is not directly caused by class imbalances, but rather, that class imbalances may yield small disjuncts which, in turn, will cause degradation
- The resampling strategy proposed by consists of clustering the training data of each class (separately) and performing random oversampling cluster by cluster
- Class-imbalance in multi-class problems
- Class-imbalance in multi-label problems

Conclusions

- Practically, it is often reported that cost-sensitive learning outperforms random resampling
- Clever re-sampling and combination methods can do quite more than cost-sensitive learning as they can provide new information or eliminate redundant information for the learning algorithm
- The relationship between training set size and improper classification performance for imbalanced data sets seems to be that on small imbalanced data sets the minority class is poorly represented by an excessively reduced number of examples that might not be sufficient for learning, especially when a large degree of class overlapping exists and the class is further divided into subclusters
- For larger data sets, the effect of these complicating factors seems to be reduced, as the minority class is better represented by a larger number of examples

Useful links

Classification with Imbalanced Datasets

References I

Barandela, R., J. L. Sánchez, V. García, and E. Rangel (2003). Strategies for learning in class imbalance problems. *Pattern Recognition 36*, 849–851.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition 30*, 1145–1159.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Chawla, N. V., A. Lazarevic, L. O. Hall, and K. W. Bowyer (2003). Smoteboost: Improving prediction of the minority class in boosting. In N. Lavrač, D. Gamberger, L. Todorovski, and H. Blockeel (Eds.), *Knowledge Discovery in Databases: PKDD 2003*, Volume 2838 of *Lecture Notes in Computer Science*, pp. 107–119. Springer, Berlin, Heidelberg.

References II

Domingos, P. (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*, KDD '99, New York, NY, USA, pp. 155–164. Association for Computing Machinery.

Drummond, C. and R. C. Holte (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II, ICML*, pp. 1–8.

Grzymala-Busse, J. W., L. K. Goodwin, and X. Zhang (2003). Increasing sensitivity of preterm birth by changing rule strengths. *Pattern Recognition Letters 24*, 903–910.

Japkowicz, N. and S.Shaju (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis 6*(5), 429–450.

Jo, D. T. and N. Japkowicz (2004). Class imbalances versus small disjuncts. *SIGKDD Explorations 6*, 40–46.

References III

Joshi, M., V. Kumar, and R. Agarwal (2001). Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In *First IEEE International Conference on Data Mining*, San Jose, CA, USA, pp. 257–264.

Kubat, M. and S. Matwin (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186. Morgan Kaufmann.

Prati, R. C., G. E. A. P. A. Batista, and M. C. Monard (2011). A survey on graphical methodsfor classification predictive performance evaluation. *IEEE Transactions on Knowledge andData Engineering 23*, 1601–1618.

Raskutti, B. and A. Kowalczyk (2004). Extreme re-balancing for svms: A case study. *SIGKDD Explorations Newsletters 6*(1), 60–69.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics–-Part B: Cybernetics SMC-6*, 769–772.

References IV

Visa, S. and A. Ralescu (2005). Issues in mining imbalanced data sets - a review paper. In *Proc. 16th Midwest Artificial Intelligence and Cognitive Science Conference*, pp. 1–7.

Zheng, Z., X. Wu, and R. Srihari (2004, 06). Term selection for text categorization on imbalanced data. *SIGKDD Explorations 6*, 80–89.