Semi-Supervised Self-training Approaches for Imbalanced Splice Site Datasets

Ana Stanescu and Doina Caragea
Department of Computing and Information Sciences, Kansas State University
Manhattan, KS 66506, USA
(anas, dcaragea)@ksu.edu

Abstract

Machine Learning algorithms produce accurate classifiers when trained on large, balanced datasets. However, it is generally expensive to acquire labeled data, while unlabeled data is available in much larger amounts. A cost-effective alternative is to use Semi-Supervised Learning, which uses unlabeled data to improve supervised classifiers. Furthermore, for many practical problems, data often exhibits imbalanced class distributions and learning becomes more challenging for both supervised and semi-supervised learning scenarios. While the problem of supervised learning from imbalanced data has been extensively studied, it has not been studied much for semi-supervised learning. Thus, in this study, we carry out an empirical evaluation of a semi-supervised learning algorithm, specifically self-training based on Naïve Bayes Multinomial (NBM), and address the issue of imbalanced class distributions both at data-level (by re-sampling) and algorithmic-level (using cost-sensitive learning and ensembles). We conduct our study on the problem of splice site prediction, a problem for which the ratio of positive to negative examples is very high. Our experiments on five different datasets show that a simple method that adds only positive instances to the labeled data in the semi-supervised iterations produces consistently better results when compared with other methods that deal with data imbalance.

1 Introduction

Quality data in adequate amounts is critical for building successful prediction models, but often times, in practice, one class (usually the more interesting class) is underrepresented. The class imbalance phenomenon occurs when the minority class is either very difficult to acquire or when the minority examples are indeed atypical relative to other cases. Generally, anomaly or novelty detection problems exhibit highly imbalanced data. Some specific applications include credit card frauds, cyber intrusions, medical diagnosis, face recognition, detecting defects in error-prone software modules, etc. Having a major unevenness between the prior class probabilities leads to impartial learning that severely alters the performance of classifiers which would otherwise give acceptable results. For supervised learning, many solutions have been proposed both at data level (under-sampling and over-sampling), as well as at the algorithm level (cost-sensitive learning and ensemble methods); for a comprehensive survey, the reader is directed to [4].

At data level, re-sampling techniques are some of the most natural and easy remedies that can be used to adjust the class distribution. The instance selection can be randomized or subject to more informative criteria [7, 1]. Learning from under-sampled data is susceptible to information loss, but can speed up the learning process as the size of the data is substantially decreased. On the other hand, over-sampling may lead to longer computation time and over-fitting because of instance duplication. Other solutions, more algorithmic-oriented, involve cost-sensitive learning [11], active learning [10], injecting extra knowledge and maybe even human interaction during the learning process. While not specifically designed for imbalanced class distributions, collections of learners usually produce better performance than a single individual classifier. Ensemble of classifiers using bagging, boosting and hybrid-approaches for imbalanced datasets were reviewed by Galar et al. [3] in the supervised framework.

In many domains, obtaining labeled data is an expensive process that requires time and human expertise. For example, in the biological field, where massive amounts of DNA data are generated thanks to cost-effective Next Generation Sequencing technologies, wet-lab analysis remains expensive and tedious. One of the most appealing ways to avoid the cost of having experts manually label data is via automated semi-supervised learning (SSL), which uses both labeled and unlabeled data in training, typically small amounts of labeled data along with much larger volumes of unlabeled data. Examples of bioinformatics problems successfully addressed with SSL approaches include alternative splicing prediction [18, 19], disease genes detection [14], prediction of cancer recurrence based on gene expression [16], etc. Furthermore, the work in [8, 21] has shown the usefulness of explicitly addressing the class imbalance problem for protein classification. For DNA classification, the class imbalance problem has been addressed in the supervised framework [20]. Specifically in this work, Wei et al. [20] study the classification of human missense phenotype prediction problem, using Support Vector Machines (SVM) in a supervised scenario. However, a systematic study of how SSL algorithms behave for imbalanced DNA classification problems has not been performed. This is precisely the problem that we address, which is to study the effect of class imbalance in a semi-supervised learning framework.

To gain a better understanding of the behavior of SSL algorithms for highly skewed DNA data, we base our study on splice site prediction using self-training [22]. Self-training is one of the most popular SSL algorithms, along with Expectation Maximization (EM), co-training, transductive SVM, and graph-based methods. Self-training can be seen as a simple wrapper method applied to a base classifier.

Splice sites are intron-exon junctions. They can be seen as relevant signals for the alternative splicing process, which regulates transcription and ultimately gene expression. We use five large DNA datasets with a positive to negative ratio of 1 to 99. Starting with these datasets, we disregard some of the instances in order to reach milder levels of class imbalance, and then gradually increase the level to study how the performance varies with the ratio. Although there are other methods to identify splice sites, this study is important at least from a theoretical perspective as it can provide useful insights for other DNA classification problems. To the best of our knowledge, this is the first attempt to systematically study how the class ratio, in the SSL framework, influences conventional solutions such as re-sampling, ensembles of self-training classifiers and cost-sensitive self-training approaches, when large, highly imbalanced DNA splice site datasets are used.

Our aim is not to get the best possible results for the splice site prediction problem, which has already been successfully addressed, among others, by Sonnenburg et al. [17] using SVM and specialized kernels, but rather to study the effects of imbalanced data on SSL algorithms. Thus, we cannot directly compare our results with the results such as those reported in [17] as both the problem addressed and the approach (supervised versus semi-supervised; SVM versus NBM) are different.

The rest of this paper is organized as follows: Section 2 describes the approaches studied. We explain how we designed our experiments in Section 3. Specifically, the data used and the feature representation are described in Section 3.1, our research questions are enumerated in Section 3.2 and the metrics used in Section 3.3. Experimental results and discussions can be found in Section 4. In Section 5 we contrast our study with other related studies. We draw some conclusions and propose future research directions in Section 6.

2 Approaches

Self-training is a bootstrapping technique in which, first, a base learner is trained on just the labeled data. Next, a randomly chosen sample from the unlabeled pool is labeled using the classifier trained on just the labeled data. From these newly labeled instances, the most confidently classified examples are added to the labeled set and the classifier retrains itself on this augmented labeled set. The process is iterative, and at each step more unlabeled instances are classified and then used in retraining. One general constraint is to maintain the positive to negative ratio of the labeled data. For example, if the class ratio in the labeled dataset is 1 to 5, then 6 examples are extracted form the unlabeled pool and added to the labeled seed set: the topmost confident positive prediction along with the top 5 most confident negative predictions. We refer to this algorithm as *self-training with imbalanced data* **(STI)** because there is no modification made to take into account the class distribution. At each iteration, the most confidently labeled examples are added, such that the original class distribution in the labeled set is maintained. The process continues until the unlabeled instances are exhausted.

We next discuss approaches that are designed to address the imbalanced data problem. The self-training approach described above does not specifically deal with this problem. Given that for our problem, there are highly skewed datasets, where the positive class can represent as little as 1% of the total number of examples, it is important to investigate the strengths and limitations of some of the most popular re-sampling techniques. As mentioned above, there are two categories of approaches for dealing with imbalanced datasets: data-level approaches and algorithmic-level approaches. In the first category, re-sampling helps to readjust the class distribution so that the learner has an equal chance of learning the positive and negative classes. Under this assumption, the labeled data can be balanced in two ways. First, under-sampling can be performed. We keep all positive instances and randomly pick negative instances until a balanced dataset is obtained. We name this variant *self-training with under-sampling* **(STU)**. Second, the minority class can be over-sampled until an equal proportion is reached. We use the Synthetic Minority Over-sampling Technique (SMOTE) proposed in [1], which is an informed technique, as opposed to a random one. In SMOTE, instances of interest are generated by interpolating other positive instances, in the feature space. They are relatively "novel" examples, whereas in random over-sampling, positive instances are simply duplicated. We opted for SMOTE because random over-sampling may increase the possibility of overfitting, since exact copies of the minority class add no new information to the dataset. We named this variation *self-training with over-sampling* **(STO)**. For these two variants, the labeled data that the learner is initially trained on is balanced (via under- or over-sampling), therefore only two instances are added into the labeled set at each iteration, the top most confident from each class.

Along the same lines, we propose a new and simple approach of dealing with the imbalance problem, which is to add only those instances that are found to be positive by the base learner. This modifies the class distribution in the semi-supervised step, but the base classifier initially trains on the labeled set which is imbalanced. We give the classifier the opportunity to see more examples from the minority class in subsequent iterations, and if an unlabeled example is classified as negative, it gets assigned a null weight. We name this new variant *self-training with positive* **(STP)**.

In the second category of approaches for dealing with imbalanced datasets, *i.e.*, algorithmic-level methods, we first use a cost sensitive approach with self-training and denote this variant as *self-training with costs* **(STC)**. Since the positive instances are so rare, they are given a higher misclassification cost. The values are equal to the imbalance coefficients, and false positives will be more penalized than false negatives. For example, if the positive to negative ratio is 1 to 99, we assign a cost of 99 for a positive instance that is classified as negative, and a cost of 1 for any negative instance that is incorrectly classified as positive.

Finally, we also investigate a variant of self-training with an ensemble approach, self-training with ensemble **(STE)**. Previous studies show that bagging several weak classifiers self-trained on bootstrapped subsamples of the labeled data outperforms multi-view training [13]. Instead of using random bootstrap sampling, our sub-classifiers are trained on balanced subspaces created using the approach from [12], to account for the imbalance problem. Specifically, each balanced subset contains all the minority instances and an equal number of majority instances sampled at random without replacement. In other words, all subsets contain the same minority instances but non-overlapping majority instances. From all these balanced subset, we learn classifiers that vote on the instances to be added to the training data at the next self-training iteration. That means, the instances with the highest averaged prediction are added back to the labeled subset of each classifier and each of them is retrained.

For all the variants described above, we use Naïve Bayes Multinomial (NBM) as the base learner for self-training. NBM is a desirable classifier in bootstrapping approaches, given that it allows for faster computation as compared to other approaches such as SVM.

3 Experimental Setup

We start this section by describing the data used in our study and the feature representation. As mentioned above, we investigate the behavior of self-training NBM variants in the context of imbalanced data, with application to the binary classification problem of predicting splice sites in a DNA sequence.

3.1 Data and Feature Representation

The acceptor splice site datasets that we used in our work were first introduced in a domain adaptation study [15]. The datasets belong to five different organisms, *C. elegans, C. remanei, P. pacificus, D. melanogaster*, and *A. thaliana*. Each instance represents a DNA sequence that is 141 nucleotides long. The AG dimer, signaling the acceptor, is set at a fixed position in the sequence, specifically the 61st. The class label indicates whether the dimer is a true acceptor site (positive class) or not (negative class). On average, each data sets contains about 160K instances, except for *C. elegans*, which contains roughly 100K instances; approximately 1% of the instances are positive. We will be representing the instances following the approach from [5]. Each sequence will have 141 features corresponding to positions in the sequence, and each feature can take one of the four values {A, C, G, T}. The value of a feature in a sequence indicates the nucleotide found at that position, corresponding to that feature.

3.2 Research Questions

Our experimental design specifically addresses the following research questions: (1) What is the most effective learning strategy when training classifiers on highly imbalanced splice site datasets in a semi-supervised framework? (2) How does the performance of the algorithms vary with the class distribution ratio?

To address the first question, we compare the variants described in the previous section, including the classical

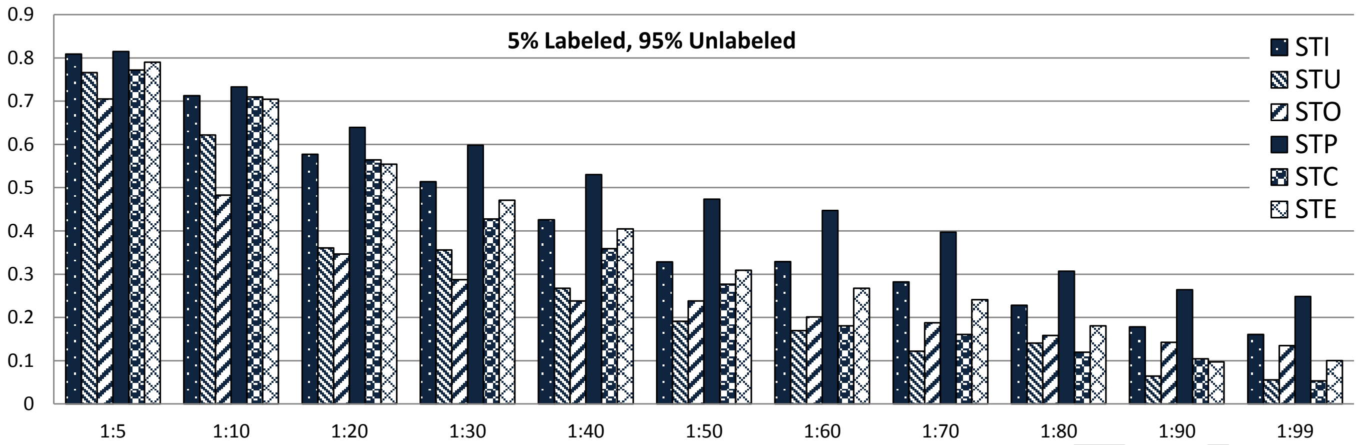

Figure 1: Averages of the auPRC values for the minority class over 5 organisms, when learning from 5% labeled data and 95% unlabeled data, while varying the positive to negative ratio from 1:5 to 1:99.

self-training approach with imbalanced data (STI), which is our baseline. The behavior of semi-supervised classifiers trained on imbalanced datasets is expected to vary with the class distribution ratio. Our second question is meant to test this hypothesis. We vary the ratio of positive to negative examples from 1:5 to 1:99. This is done by discarding some of the majority instances at random. Since semi-supervised learning is advantageous when there are far more unlabeled examples than labeled examples, we consider 5% of the data to be labeled and the remaining 95% will be treated as unlabeled. Both labeled and "unlabeled" sets are chosen at random, without replacement, from the original dataset. To simulate the unlabeled set, we simply ignore the labels of the instances in that set.

3.3 Evaluation Metrics

We evaluate our classifiers using the area under the Precision-Recall Curve (auPRC), which is a better assessment metric as compared to the area under the Receiver-Operating Curve (auROC) when tackling problems with highly imbalanced datasets [2]. Since the minority class of true acceptor splice sites is of interest, we concentrate on this class and how the algorithms can identify these positive instances.

To account for sampling variation, for each organism we perform 10-fold cross validation and average the auPRC values over the 10 folds. At each round, 90% of the data was used in training and the remaining 10% was used as test. From the 90% training data, 5% was set aside as labeled while the rest was used as unlabeled. In our graphs, we report the average (due to space limitations and the fact that the results were generally consistent) over all five organisms of the auPRC values for the positive class.

4 Results

We have summarized our results in Figure 1. The graph represents averaged auPRC values over all five organisms, obtained from classifiers trained on 5% labeled data and 95% unlabeled data. As can be seen from the figure, STP outperforms all the other models. This suggests that gradually balancing the labeled data during the semi-supervised step is a useful technique to deal with imbalanced distributions. Surprisingly, the classical approach, STI is the second best.

It has been reported that under-sampling is more suitable for semi-supervised learning on imbalanced datasets than over-sampling [9]. We have observed the same trend when the class ratio is relatively low. However, for highly imbalanced datasets (over 1:50), over-sampling seems to be a better approach.

Ensemble learning outperforms the re-sampling techniques except for the highest imbalanced cases (1:90 and 1:99). This result is consistent with the conclusions in the review by Galar et al. [3], who showed that ensemble classifiers are more effective than single classifiers trained on re-sampled data in supervised frameworks.

Cost-sensitive learning has also been shown to outperform re-sampling techniques in supervised learning [6], and the trend is maintained in the case of semi-supervised learning, for lower degrees of imbalance (up to 1:60).

As expected, overall, better values for auPRC are obtained when the class distribution ratio is smaller and they decrease for the more highly imbalanced cases because the datasets increase in size and the prediction problem becomes more difficult as the positive class is more and more underrepresented.

Over-sampling tends to perform best when the class ratio is higher (over 1:50), which shows that synthetic generation of instances is useful when there is not enough labeled data available.

To conclude, the best results were achieved by self-training with positive instances (STP), followed by the standard self-training approach with imbalanced data (STI), and then the ensemble variant (STE), closely follwed by the cost-sensitive approach (STC). However, for the most extreme case of imbalance (1:99), over-sampling and ensembles are better suited.

5 Related Work

Previous studies of semi-supervised learning from imbalanced data have focused on datasets with relatively low imbalance degrees. An interesting ensemble-based approach for the sentiment analysis problem was proposed by Li et al. [9]. Co-training was used as a base-classifier in their work. The authors experimented with four different domains, and the class ratio ranged from roughly 1:3 to 1:8. To address the class imbalance, they first created balanced subsets using the approach from [12] (also used in our study for the STE variant). Next, they dynamically generated two random feature subspaces from each subset and used co-training to learn from these subsets. Their approach showed improvement over under-sampling methods in the context of sentiment classification. Our results for STE are consistent with their findings for smaller class ratios.

One other study that deals with imbalanced data and involves semi-supervised learning was proposed by Kundu et al. [8] to address the problem of predicting SH2-peptide interactions. However, the use of semi-supervised learning in this approach is different from the use in the previously reviewed work and also from our work, as will be explained in what follows. In [8], the positive (minority) class consists of SH2-peptide interactions, and the negative (majority) class consists of non-interactions. Positive instances (interactions) can be reliably identified from high density peptide and microarray experiments. However, negative instances (non-interactions) are sometimes harder to established - lack of current evidence for an interaction could imply a non-interaction or an interaction to be discovered in the future. This is why, in some of the 51 datasets used in [8] (the largest having 400 instances), there could be up to 15 times more positive instances as compared to negative instances (a ratio that may seem counterintuitive given that the positive class represents the minority). When this happens, self-training is used to iteratively learn a model from a small, reliable dataset, followed by the use of the resulting model to identify non-interactions (unlabeled instances confidently predicted as negatives). As a result, a balanced dataset is obtained. When there are more negative instances as compared to positives, in the original training set, over-sampling is used to generate more positive instances. The final models, polynomial kernel SVMs, are trained on the resulting balanced datasets and customized through parameter validation. The results outperform state-of-the-art SH2-peptide interaction prediction tools. In conclusion, the approach in [8] introduces an alternative usage for self-training, that might be useful for many biological classification problems, where the negative/majority class cannot always be established reliably. This includes also our splice site prediction problem, as it can happen that for some of the sites considered to be negative in our data, there could be later evidence to show that they are, in fact, positives.

6 Conclusions and Future Work

In this study, we have performed an analysis of self-training classifiers on imbalanced data. Empirical evidence on five large DNA datasets shows that a simple self-training variation (STP) that balances the labeled sets with only instances classified as positive, can consistently exceed other standard methods by as much as 7% in most cases. Self-training on the original imbalanced sets (STI) was the second best variant. The next best performance came from the ensemble (STE) and cost (STC) variants for most of the cases, with a notable exception for the over-sampling variant (STO), which performed better for higher degrees of class imbalance (1:90 and 1:99).

As part of future work, experimenting with different DNA datasets could offer additional insight into the problem. Co- or multi-training as well as representing the instances with 2 or more complementary views might potentially increase the overall performance. Transductive approaches represent another avenue for future work. At last, we would like to experiment with approaches like the one introduced in [8] in order to identify possible examples in our DNA dataset mislabeled as negatives.

Acknowledgements

The computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants ACI-1341026, CNS-1126709, CNS-1006860, and EPS-0919443.

References

[1] N.V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, 2002.

[2] J. Davis, Jesse and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves", *Proceedings of the 23rd International Conference on Machine Learning*, ICML, p. 233-240, 2006.

[3] M. Galar, A. Fernndez, E. Barrenechea, H. Bustince, F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches" *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Apps. and Reviews*, Vol. 42, No. 4, p.463-484, 2012.

[4] H. Haibo and E. A. Garcia, "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Eng*, Vol. 21, Issue 9, p. 1263-1284, 2009.

[5] N.Herndon, D. Caragea, "Empirical study of domain adaptation with Naïve Bayes on the task of splice site prediction", *Proc. of the 5th Intl. Conf. on Bioinformatics Models, Methods and Algorithms*, BIOINFORMATICS, 2014.

[6] N. Japkowicz, S. Shaju, "The class imbalance problem: A systematic study" *Intelligent data analysis* Vol. 6, No. 5, p. 429-449, 2002.

[7] J.N. Korecki, R.E. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer, "Semi-supervised learning on large complex simulations", *19th Intl Conf. on Pattern Recognition*, ICPR, p. 1-4, 2008.

[8] K. Kundu, F. Costa, M. Huber, M. Reth, R. Backofen, "Semi-supervised prediction of SH2-peptide interactions from imbalanced high-throughput data", *PLoS ONE, Public Library of Science*, Vol. 8, No. 5, 2013.

[9] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification", *Proc. of the 22nd Intl Joint Conf. on Artificial Intelligence*, Vol. 3, 2011.

[10] S. Li, J. Shengfeng, Z. Guodong, and L. Xiaojun, "Active learning for imbalanced sentiment classification", *Proc. of the Joint Conf. on Empirical Methods in Nat. Lang. Processing and Computational Nat. Lang. Learning*, ACL pp. 139-148, 2012.

[11] C. X. Ling , and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem", *Encyclopedia of Machine Learning*, 2008.

[12] X.Y. Liu, J. Wu, Z.H. Zhou, "Exploratory undersampling for class-imbalance learning", *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, Vol. 39, No. 2, p.539-550, 2009.

[13] V. Ng, and C. Cardie, "Weakly supervised natural language learning without redundant views" *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics on Human Language Technology*, Vol. 1, 2003.

[14] T.P. Nguyen and Ho Tu-Bao. "Detecting disease genes based on semi-supervised learning and proteinprotein interaction networks" *Artificial Intelligence in Medicine*, Vol. 54, No. 1 p 63-71, 2012.

[15] G. Schweikert, G. Rätsch, C. K. Widmer, and B. Schölkopf, "An Empirical Analysis of Domain Adaptation Algorithms for Genomic Sequence Analysis", *Advances in Neural Information Processing Systems*, p. 1433-1440, 2008.

[16] M. Shi, B. Zhang "Semi-supervised learning improves gene expression-based prediction of cancer recurrence", *Oxford Journal Bioinformatics*, 2011.

[17] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, G. Rätsch, "Accurate splice site prediction using support vector machines", *BMC Bioinformatics*, Vol.8, No.10, p.1-16, 2007.

[18] A. Stanescu, D. Caragea, "Semi-supervised learning of alternatively spliced exons using Expectation Maximization type approaches", *Proc. of the 3rd Intl. Conf. on Bioinformatics Models, Methods and Algorithms*, BIOINFORMATICS, 2012.

[19] K. Tangirala, D. Caragea, "Semi-supervised learning of alternative splicing events using Co-Training", *Proc. of the IEEE International Conference on Bioinformatics and Biomedicine*, 2011.

[20] Q. Wei, and R. L. Dunbrack Jr., "The role of balanced training and testing data sets for binary classifiers in bioinformatics", *PLoS ONE, Public Library of Science*, Vol. 8, No. 7, 2013.

[21] Q. Yanjun, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins." *Bioinformatics*, Vol. 26, No. 18, p. 645-652, 2010.

[22] D. Yarowsky. "Unsupervised word sense disambiguation rivaling supervised methods". *Proc. of the 33rd annual meeting on Association for Computational Linguistics*, ACL, 1995.