Ensemble-based semi-supervised learning approaches for imbalanced splice site datasets

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

***Abstract*—Producing accurate classifiers depends on the quality and quantity of labeled data. The lack of labeled data, due to its expensive generation, critically affects the application of machine learning algorithms to biological problems. However, unlabeled data may be acquired relatively faster and in larger quantities thanks to current biochemical technologies, called Next Generation Sequencing. In such cases, when the number of labeled instances is overwhelmed by the number of unlabeled instances, semi-supervised learning represents a cost-effective alternative that can improve supervised classifiers by utilizing unlabeled data. In practice, data oftentimes exhibits imbalanced class distributions, which represents an obstacle for both supervised and semi-supervised learning. The problem of supervised learning from imbalanced datasets has been extensively studied, and various solutions have been proposed to produce classifiers with optimal performance on highly skewed class distributions. In the case of semi-supervised learning, there are not as many efforts aimed at the imbalance data problem. In this paper, we study several ensemble-based semi-supervised learning approaches for predicting splice sites, a problem for which the imbalance ratio is very high. We run experiments on five imbalanced datasets with the goal of identifying which variants are the most effective.**

***Index Terms*—semi-supervised learning; imbalanced datasets; ensemble; self-training**

I. Introduction

In domains such as online social media, biology, or medicine, the research challenge has shifted from producing data to interpreting data. For genetics, the bottleneck lies in the interpretation and labeling of massive amounts of raw DNA data produced by next generation sequencing technologies. Machine learning and statistical analysis are practical and efficient ways of analyzing and interpreting data. Supervised learning is an effective technique that can assist in the annotation process, but supervised learning algorithms require large labeled datasets in order to produce useful classification systems. In any domain, the process of labeling data is an expensive task and for biology in particular, wet-lab experiments remain costly and tedious, as they require human expertise and time. If having experts manually label more data is not an option (due to high costs), a desirable alternative is to leverage the much larger quantities of unlabeled data (when available). This automated approach, namely semi-supervised learning (SSL), typically utilizes small amounts of labeled data and considerably larger amounts of unlabeled data in training, with the ideal goal of improving upon a classifier trained only on the labeled data.

Improving supervised classifiers by leveraging unlabeled data is a very attractive concept, yet it does not always work as intended. In practice, it is very common for a classifier to be degraded by the unlabeled data [1]. Deciding whether or not to use the unlabeled data is a problematic task [2], and the focus of ongoing research [3]. Moreover, imbalanced datasets pose serious problems for all classifiers, both supervised and semi-supervised (transductive also), but the supervised learning field enjoys a richer collection of solutions to overcome problems raised by imbalanced class distributions. Ideally, datasets should be sufficiently large, as well as balanced in order to result in classifiers that learn significantly better than a random model; but many times in practice, one common obstacle any learning algorithm must overcome is the class imbalance problem. This is a phenomenon that occurs when examples from one class, usually the class of interest, are very difficult to acquire or are genuinely atypical, in comparison with the other classes. For example, novelty or anomaly detection problems are affected by high imbalance. More specific applications include credit card frauds, cyber intrusions, medical diagnosis, face recognition, detecting defects in error-prone software modules, etc. A significant disproportion in the class prior probabilities usually leads to biased learning.

Classifiers that otherwise behave favorably (in the presence of balanced, or mildly imbalanced data), are negatively affected when learning from non-uniform distributions. In the supervised context, many solutions have been proposed. At data level, under-sampling is the most straightforward solution. Adjusting the class distribution can be done by simply discarding instances from the majority class. In this case, the trade-off is between information loss and speed of learning. Selecting which instances to keep (or conversely, discard) can be done randomly, or in a more informative way [4]. Another case of re-sampling is over-sampling. The caveats here include longer computation times and over-fitting caused by instance replication. This process can also be randomized, although not much information is gained - since duplicate instances can be viewed as one instance with increased weight. Creating completely new artificial instances represents a more informative way of over-sampling, *e.g.*, the SMOTE technique [5]. At algorithmic level, solutions involving cost matrices are most common. The cost matrix contains penalties associated with different classification errors (as some mistakes are more serious than others). Other techniques include active learning [6], injecting

extra knowledge and maybe even human interaction during the learning process. Ensemble learning is another approach, even though it was not originally intended for imbalanced distributions. Collections of classifiers using bagging, boosting and hybrid approaches specifically targeting imbalanced data were reviewed by Galar *et al.* [7] in the supervised framework.

For SSL, there are notable studies that explore the imbalanced problem and propose effective solutions, but the imbalance degrees are moderate (up to 1-to-40). Our aim in this study is to adapt existing solutions to datasets with higher degrees of imbalance (up to 1-to-99), and study their behavior on significantly disproportionate class distributions when the labeled data available is less than 1%. Such a small amount of labeled data is expected to lead to weak classifiers, but an ensemble of classifiers can help overcome, to some extent, this shortcoming. It has been shown by Galar *et al.* [7] that ensembles perform better than single learners trained on resampled data in supervised frameworks. Another study by Li *et al.* [8] led to the same conclusion, that an ensembles of co-training algorithms is suitable for imbalanced datasets.

To better understand the behavior of SSL algorithms for DNA prediction problems when the data suffers from severe class asymmetry, we explore the problem of classifying splice sites. We are interested in observing *if* and *how* small amounts of labeled data influence the behavior of semi-supervised algorithms. We use ensembles of self-trained classifier and address the imbalance issue in three ways - (1) by creating balanced subsets to train the initial sub-classifiers of the ensemble, (2) by dynamically balancing the total amount of labeled data during the semi-supervised iterations, and (3) by ensuring the sub-classifiers remain diverse enough, such that the ensemble benefits from their bagged voting. Specifically, we explore SSL algorithms based on ensembles of classifiers in the context of acceptor splice site prediction. Splice sites are found at the boundaries between intron-exon junctions (in the case of acceptor splice sites) and between exon-intron junctions (in the case of donor splice sites). They are relevant signals to the alternative splicing process thereby regulating transcription and gene expression. Generally, splice sites are canonical, which means they are indicated by the presence of the dimers "AG" and "GT" for acceptor and donor sites, respectively. However, the simple occurrence of the dimer is not enough to declare a splice site, as the possibilities are enormous for a 2-nucleotide-long sequence to appear in a genome. Fortunately, the areas surrounding splice sites exhibit strong consensus sequences which, although somewhat different from one organism to another, can help statistical analysis and prediction algorithms. Splice site prediction consists of two extremely imbalanced classification tasks: discriminating between true acceptor sites and decoy positions with the AG dimer (which is the problem we are addressing in this paper) and discriminating between true donor sites and decoy positions, with the GT dimer.

The rest of the paper is organized as follows: we review similar work in Section II and present the context and need for our study. The methods we used are described in Section III. In Section IV we describe the data, the research questions and our experimental setup. We discuss our results in Section V and present our conclusions inSection VI, where we also enumerate several directions we are interested in pursuing as future work.

II. Related Work

Semi-supervised learning was successfully applied to many bioinformatics problems, including predicting alternatively spliced exons [9], [10], detecting disease genes [11], predicting cancer recurrence based on gene expression [12], classifying protein domains into SCP (Structural Classification of Proteins) super-families [13], predicting protein localization [14], [15], motif discovery [16], and also in problems related to gene regulatory networks [17], [18].

For many applications in bioinformatics, the imbalance data problem is also prevalent; Qi *et al.* [19] and Kundu *et al.* [20] have shown the usefulness of explicitly addressing the class imbalance problem for protein classification. Qi *et al.* [19] enhance semi-supervised multi-task learning by using auxiliary information and successfully detect interacting pairs between human proteins and HIV-1 proteins. For this problem, the imbalance is caused by the fact that truly interactive protein pairs are rare, yet there are numerous examples of protein pairs which could potentially interact (but currently lack experimental proof). Kundu *et al.* [20] use SSL (self-training in particular) to first balance a dataset of SH2-peptide interactions, where the positive class constitutes interactions, and the negative class consists of non-interactions. The negative class of non-interactions is more difficult to establish, because the simple lack of current evidence does not necessarily mean the interaction cannot occur in future circumstances. In their case, self-training is used to identify non-interactions. Having now extra confidence that the instances are indeed negative, a final supervised model is trained on the balanced datasets, namely Support Vector Machines (SVM) with a polynomial kernel. This innovative usage of SSL could benefit a larger collection of biological classification problems where one class cannot always be reliably determined. The splice site prediction problem falls in the same category because some splice sites which are currently established as negative, might prove later on to be, in fact, positives.

Kondratovich *et al.* [21] used Transductive Support Vector Machines (TSVM) on small but imbalanced datasets for the problem of molecule activity prediction. Experiments on 10 datasets with imbalance ratios of up to 1-to-40 and a maximum of 3,000 instances demonstrated the effectiveness of TSVMs in overcoming obstacles posed by imbalanced data.

Li *et al.* [8] found a solution for SSL from imbalanced data in the domain of sentiment classification. Their approach employs an ensemble of co-trained classifiers. Each sub-classifier is learned from a balanced subset (obtained using a technique initially recommended by Liu *et al.* [22] - which we also used in this study, and describe in Section III). For co-training, Li *et al.* [8] dynamically generated two views by randomly sub-sampling the feature space. The authors experimented with four different domains, and the class ratio

of the datasets ranged from roughly 1-to-3 to 1-to-8. Co-training requires the data to be represented according to two views, where each view should be sufficient for classification and independent of the other view given the class [23], the objective being to learn classifiers that inform each other about their best predictions on the unlabeled data. Since our datasets exhibit imbalance degrees of 1-to-99, having two views for each of the 99 sub-classifiers would unnecessarily increase the memory complexity - a trade-off we did not find worthy at the moment. As randomly generated views do not necessarily satisfy these conditions, in our approach we use self-training to avoid any possible problems caused by random splits and the need for more sophisticated ways to split the features.

Korecki *et al.* [4] also used self-training and ensembles of random forests on a multi-class problem to discriminate between complex simulations. In the semi-supervised step, they used a more informative approach of deciding which instances to add to the labeled set: a threshold based on the average Euclidean distances to the centroids found in the labeled set.

For DNA classification, the class imbalance problem has been addressed in the supervised framework by Wei *et al.* [24]. Specifically, the authors study the classification of human missense phenotype prediction problem, using SVM in a supervised scenario.

In our previous work [25], we studied a variety of techniques to alleviate the imbalance data issue in the semi-supervised framework for the DNA prediction problem of identifying acceptor splice sites. We found that dynamically balancing the labeled dataset during the semi-supervised self-training iterations was the most successful, surpassing re-sampling techniques (random under- and SMOTE over-sampling), cost-sensitive and ensemble approaches. In this paper, our aim is to experiment with ensembles of self-training classifiers that bootstrap the unlabeled data and dynamically balance the initial labeled set, when very small amounts of labeled data are available (less than 1% of the total data, including unlabeled). We compare our ensemble-based variants (where sub-classifiers are trained on balanced subsets of the labeled data) with a state-of-the-art approach from the domain of sentiment classification [8], which we adapted for self-training. Splice sites can be accurately identified using SVM and specialized kernels, as Sonnenburg *et al.* have shown in [26]. As opposed to Sonnenburg *et al.* who used supervised SVMs, we are addressing the SSL case using Naïve Bayes classifiers. A direct comparison is not our primary objective, nor possible, since the problem addressed and the approach (supervised versus semi-supervised; SVM versus Naïve Bayes) are different.

III. Methods

In this section we describe the types of methods we are studying. Our focus is on ensemble methods and what variations of semi-supervised ensembles produce the best results.

We use self-training, an algorithm introduced by Yarowsky [27] for a natural language processing problem. Self-training is a very popular semi-supervised algorithm, together with Expectation Maximization (EM), co-training, transductive Support Vector Machines, and graph-based methods. Self-training is a simple wrapper method that can make use of any base classifier. First, the base classifier is trained on the labeled data and then used to classify the unlabeled data. The newly labeled instances are subsequently used to self-train in the next iteration, by integrating them in the labeled set and re-training the classifier. An important requirement is to maintain the ratio of positive to negative instances in the labeled training set when adding the newly labeled instances. Self-training is an iterative procedure whose goal is to enlarge the labeled dataset by accepting its own predictions and incorporating them as labeled data in order to ultimately produce a better model.

Combining the predictions from multiple diverse classifiers produces more accurate results than any single classifier from the ensemble, especially when the individual classifiers are weak (slightly better than random guessing) [28]. Previous studies have shown that bagging several weak classifiers self-trained on bootstrapped subsamples of labeled data outperformed multi-view training [29]. In our work, we train each individual classifier on a much smaller but balanced subsample of the labeled data instead of utilizing random bootstrap sampling. Specifically, we create balanced subsets from the labeled data by replicating all the minority instances and under-sampling without replacement the majority instances, an approach introduced by Liu *et al.* [22]. We use as few labeled instances as possible to still be able to create balanced subsets in each case of imbalance degree, while maintaining the total amount of labeled data under 1%. Each balanced subset is used to train a base classifier (in our case, Naïve Bayes) and every classifier produces a prediction probability for each unlabeled instance. The predicted probabilities from the sub-classifiers are averaged and the decided label will be assigned to the unlabeled instance. At each self-training iteration we only label a fixed sample size (randomly picked from the unlabeled data) for efficiency purposes. From the newly labeled instances in this sample, we fetch the most confidently classified ones to augment the labeled dataset with. The remainder of the sample is simply discarded, and a new sample is picked for the next iteration; this procedure ensures that the entire unlabeled dataset is analyzed (classified) once, and that the instances labeled with less certainty are not allowed to compromise the classifier. The ensemble obtained at the end of this process (after the unlabeled data is exhausted), decides the labels for the test data in a similar fashion: predictions from the individual classifiers are averaged to produce the final label.

STEO (Self-Training Ensemble Original) This variant is inspired by the approach of Li *et al.* [8]. The top most confident pseudo-labeled examples are added to the labeled data of each sub-classifier. The original balanced classifiers are maintaining their subsets balanced, as an equal number of positive and negative instances augment the labeled subsets after each iteration. In their approach, Li *et al.* used just two instances (the topmost confidently labeled positive example and the topmost confidently labeled negative example) to

enlarge the initial labeled dataset. Then, they iterated the self-training step 50 times and observed the performance variation from one iteration to another. This differs from our approaches because instead of classifying all the unlabeled data at every iteration, we only classify a subsample of the unlabeled data, and select the best instances to retrain with, while discarding the rest.

STEM (Self-Training Ensemble Modified) This variant is a modification of the original approach STEO. We are interested in exploiting the unlabeled data as much as possible, especially the positive instances that potentially exist therein. We use an augmentation factor of 10, meaning that after each iteration, we add 10 newly-labeled instances (from the selected sample) to each class. In other words, the same 20 most confidently labeled instances (10 positives and 10 negatives) are added to the labeled subset of each self-trained Naïve Bayes sub-classifier. This will increase the labeled dataset at a faster rate, and having more pseudo-labeled instances to train from would potentially speed up the improvement in performance. Basically, STEO is STEM with an augmentation factor of 1. The rest of the ensemble variants that we use in this paper (and are described next) use the same augmentation factor of 10, which we chose in order to speed up the learning process.

STEX (Self-Training Ensemble eXtended) In this variation, each sub-classifier's labeled subset is augmented with a new pseudo-labeled set that is not balanced, but maintains the class imbalance of the original dataset. Although we start off with balanced sub-classifiers, during the semi-supervised iterations their training datasets are augmented with 1-to-N newly labeled examples (they all voted for), where N represents the imbalance degree. For example, if the dataset exhibits an imbalance ratio of 1-to-50, we train 50 sub-classifiers on balanced subsets obtained with the method proposed by Liu *et al* [22]. After a self-training phase, where the ensemble votes on new labels, each sub-classifier is augmented with 10 positive examples and 500 negative examples. Intuitively, this approach will start off diverse enough and eventually will adapt to the imbalance degree of the dataset, thus being able to ultimately capture better the distribution from the test data and achieve higher performance.

STEP (Self-Training Ensemble Positive) This approach is a hybrid between our previous approach, STP (which was specifically designed to address the imbalanced data problem by adding only positive instances to the labeled dataset during the self-training iterations [25]) and STEM (the ensemble approach of sub-classifiers trained on balanced subsets that vote to add new instances during the semi-supervised iterations, and that use an augmentation factor of 10).

STED (Self-Training Ensemble positive Distributed) This is another newly proposed approach for the imbalanced data problem in the context of semi-supervised learning of splice site datasets. It is a combination of our previous findings (the dynamic balancing [25]) and assumptions (maintaining the diversity of the sub-classifiers during the semi-supervised steps). STED is an ensemble of self-trained classifiers that augment their labeled subsets only with instances that have been voted positive and, furthermore, the newly labeled instances are distributed among the sub-classifiers such that the learners remain different enough to capture distinct (or more diverse) information, and, thus benefiting the ensemble more.

LBE (Lower Bound Ensemble) In order to compare the semi-supervised ensemble variants to a supervised lower bound, we created the corresponding ensemble supervised approach, in which the same balancing technique is used to create the subsets. Supervised Naïve Bayes classifiers are trained on the balanced subsets and the average of the probabilities from the sub-classifiers represents the verdict of the ensemble on a test instance. The unlabeled instances are not utilized at all in this supervised approach.

IV. Experimental Setup

Experiments are conducted on five imbalanced datasets from a domain adaptation study [30]. Each dataset represents DNA sequences from five organisms: *C. elegans, C. remanei, P. pacificus, D. melanogaster*, and *A. thaliana*. Each sequence contains 141 nucleotides and the dimer "AG", which signals the acceptor splice site, is fixed at position 61 in the sequence. The datasets contain approximately 160,000 instances, with the exception of the *C. elegans* dataset, which contains approximately 120,000 instances. For each organism, approximately 1% of the instances are positive (true acceptor splice sites). We used the feature vector representation from [25].

Our experimental setup is specifically designed to address the following research questions: (1) Is supervised learning aided by additional unlabeled data in the case of highly imbalanced datasets, or do the pseudo-labeled instances deteriorate the classification performance? (2) How does the performance of SSL algorithms based on ensembles of classifiers vary with the class distribution ratio? (3) What is the most effective ensemble variant when training classifiers on highly imbalanced splice site datasets in a semi-supervised framework?

To simulate an SSL environment, we followed the approach from [25]; for labeled data, we picked instances randomly and for unlabeled data, we simply ignored the labels. To answer the first question, we kept the labeled data to a minimum. We randomly picked positive and negative examples, and ensured the imbalance degree was maintained, which in general meant that labeled instances represented considerably less than 1% of the training dataset, as we wanted to use just enough labeled instances to create the balanced subsets, but not more than 1%. For the second and third questions, in order to observe how the algorithms' performance varies with the imbalance degree, we re-sampled the original datasets to simulate different class distributions. For every organism, we varied the proportion of positive to negative instances from 1-to-5 to 1-to-99. Classifiers are highly susceptible to the order in which the data arrives, especially semi-supervised learners which iterate through the unlabeled instances, therefore we built our datasets incrementally, in the sense that larger datasets (with higher data imbalance) were obtained by adding more instances to the smaller datasets (with lower data imbalance) until the dataset became the original set (with the imbalance ratio of 1-to-99).

To test the performance of our algorithms and to avoid any sampling bias, we used 10-fold cross validation. For each fold, 90% of the data was used in training (as labeled and unlabeled instances) and the remaining 10% was used for testing. From the 90% training, we picked labeled instances such that the ratio was maintained and the labeled instances represented less than 1% of the data. The cross-validation technique ensures that each instance of the dataset is tested one time, and that the test folds reflect the characteristics of the data.

Traditional evaluation metrics, such as accuracy or error rate, are not suitable to judge the performance of imbalanced learning results. For example, a learner that classifies all the test instances as the majority class on a dataset with an imbalance ratio of 1-to-99 achieves a 99% accuracy, obviously meaningless. We evaluated our classifiers' performance using the area under the Precision-Recall Curve (auPRC), a more appropriate assessment metric compared to the area under the Receiver-Operating Curve (auROC) when undertaking problems with highly imbalanced datasets [31]. Since the minority class of true acceptor splice sites is of interest, we concentrated on this class and how the algorithms can identify positive instances, therefore we used the auPRC value of the positive class to determine and compare the quality our models.

V. Results and Discussion

The results are presented in TABLE I. For easier interpretation, we report the averaged values of the auPRC for the positive class over the five organisms, as the trends are generally maintained for individual organisms. The first column represents the imbalance degree of the dataset (positive-to-negative ratios). The LBE column represents the supervised lower bound against which we compare the SSL algorithms.

Imbal	LBE	STEO	STEM	STEX	STEP	STED
1-to-5	0.431	0.4292	**0.4636**	**0.5514**	**0.535**	**0.5478**
1-to-10	0.4122	0.4	0.3508	**0.4862**	**0.4674**	**0.528**
1-to-20	0.437	0.3762	0.239	0.4356	**0.502**	**0.4692**
1-to-25	0.4406	0.3726	0.189	0.3178	0.3876	**0.4816**
1-to-30	0.4316	0.3798	0.1624	0.2942	0.4134	**0.488**
1-to-40	0.4684	0.414	0.1342	0.3058	0.3844	**0.4934**
1-to-50	0.4694	0.3556	0.1264	0.2386	0.379	**0.483**
1-to-60	0.4876	0.358	0.1468	0.2086	0.3086	0.4756
1-to-70	0.4666	0.38	0.1044	0.1766	0.334	**0.476**
1-to-75	0.4716	0.3688	0.1702	0.24	0.3308	**0.473**
1-to-80	0.465	0.3354	0.132	0.1286	0.2812	**0.4684**
1-to-90	0.4722	0.358	0.125	0.1644	0.306	0.4626
1-to-99	0.445	0.3628	0.1276	0.18	0.2852	**0.469**

TABLE I: Averages of the auPRC values for the positive class over the five organisms, when the class imbalance ratio varies from 1-to-5 to 1-to-99 and the amount of labeled instances represents less than 1%. **LBE** (Lower Bound Ensemble), **STEO** (Self-Training Ensemble Original [8]), **STEM** (Self-Training Ensemble Modified), **STEX** (Self-Training Ensemble eXtended), **STEP** (Self-Training Ensemble Positive), **STED** (Self-Training Ensemble positive Distributed). Emphasized values represent improvements over the supervised LBE.

We start our discussion of the results by answering the first research question from Section IV, which simply put was "Does unlabeled data help?" The original STEO approach always falls below the supervised lower bound but the other variants improve upon it: STEM in the case of 1-to-5, STEX in the case of 1-to-5 and 1-to-10, STEP in cases with imbalance of up to 1-to-15, and STED, which almost always outperforms the lower bound, most notably in the extreme case of 1-to-99.

The second research questions revolves around the ensemble approaches. The only two approaches that do not degrade with higher imbalance ratios and maintain a somewhat constant value of auPRC are STEO (from [8]) and our proposed approach STED. As opposed to STEO, which does not raise above the supervised lower bound, STED seems to benefit from the unlabeled data and produces higher auPRC values. Surprisingly, STEM showed the worse performance. STEM is mainly STEO but instead of augmenting the labeled set with just two instances (the most confidently labeled one positive instance and one negative instance), STEM is augmented with 20 instances (10 from each class). This behavior denotes that the initial classifiers were misguided by unlabeled data and since 10 times more misclassified instances were introduced, the performance decreased just as drastic. The same pattern was followed by STEX, which achieved a much better accuracy than STEO for datasets with milder degrees of imbalance. However, STEX recorded the most abrupt decrease - not surprising, since the extended version also incorporated more negative instances during the iterations.

The answer to the third research question question is STED, the most useful classifier in this set of experiments. Although its performance increase over the lower bound was not substantial (3%), it is still a considerable improvement over the other semi-supervised approaches (up to 10% above the next best performance, that of STEP).

As a general trend, the semi-supervised algorithms' performance degrades as the datasets become more imbalanced. For greater imbalance degrees (1-to-90, 1-to-99), even though the initial supervised classifiers have more labeled data to learn from in the early stages than the algorithms trained on datasets with lower imbalance degrees (1-to-5, 1-to-10), the performance degrades. This trend is expected because the learning is impaired by a strong unevenness in the prior distribution. On the other hand, the supervised algorithm LBE shows an increase in auPRC values as the imbalance degree increases. Generally, we observed that one of our proposed methods, namely STED, had superior performance to that of the other approaches, and was the only method to achieve increased performance over the supervised lower bound. The auPRC value for the positive class increased with 3% on average for all the imbalance degrees compared to the supervised lower bound. For all the other semi-supervised methods the unlabeled data proved to be detrimental.

VI. Conclusions

In this paper we propose several ensemble-based variants of semi-supervised learning algorithms adapted to highly imbal-

anced datasets and test their performance on five large splice site datasets. One approach, STED (Self-Training Ensemble positive Distributed), is producing the best results and its main characteristics are: (1) the sub-classifiers in the ensemble must maintain their diversity thus each adding different instances retrieved from the unlabeled data; (2) the ensemble should be dynamically balanced by only adding positive instances in the semi-supervised iterations. These hypotheses lead to a successful utilization of the unlabeled data. Our empirical results show that with less than 1% labeled data, the proposed method STED can successfully leverage the unlabeled data and produce classifiers that outperform the supervised lower bound in most cases, including the case in which the imbalance degree is maximum (1-to-99), for an overall average of 3%. Although not entirely "safe", we have shown that the proposed approach notably surpasses all the other semi-supervised variants, and can be considered a stepping stone towards further improving such semi-supervised learning methods for datasets where one class is severely underrepresented.

In future work, we consider using other types of base learners (*e.g.*, large margin classifiers) for the self-training and possibly co-training algorithms. It is also of interest to explore the behavior of SVMs in a transductive approach, as the results from Kondratovich *et al.* [21] showed great potential for the problem of protein classification from imbalanced datasets.

Acknowledgment

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] Y.-f. Li and Z.-h. Zhou, "Towards making unlabeled data never hurt," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1081–1088.

[2] A. Singh, R. Nowak, and X. Zhu, "Unlabeled data: Now it helps, now it doesn't," in *Advances in Neural Information Processing Systems*, 2009, pp. 1513–1520.

[3] Y. Wang and S. Chen, "Safety-aware semi-supervised classification," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 11, pp. 1763–1772, Nov 2013.

[4] J. N. Korecki, R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "Semi-supervised learning on large complex simulations," in *Proceedings of The Nineteenth International Conference on Pattern Recognition, ICPR 2008*. IEEE, 2008, pp. 1–4.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique"," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[6] S. Li, S. Ju, G. Zhou, and X. Li, "Active learning for imbalanced sentiment classification," in *Proceedings of The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 139–148.

[7] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and 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: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[8] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification," in *Proceedings of The Twenty Second International Joint Conference on Artificial Intelligence-Volume Volume Three*. AAAI Press, 2011, pp. 1826–1831.

[9] K. Tangirala and D. Caragea, "Semi-supervised learning of alternatively spliced exons using co-training," in *The 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2011, pp. 243–246.

[10] A. Stanescu and D. Caragea, "Semi-supervised learning of alternatively spliced exons using Expectation Maximization type approaches," in *Proceedings of The Third International Conference on Bioinformatics Models, Methods and Algorithms*, 2012, pp. 240–245.

[11] T.-P. Nguyen and T.-B. Ho, "Detecting disease genes based on semi-supervised learning and protein-protein interaction networks," *Artificial Intelligence in Medicine*, vol. 54, no. 1, pp. 63–71, 2012.

[12] M. Shi and B. Zhang, "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," *Bioinformatics*, vol. 27, no. 21, pp. 3017–3023, 2011.

[13] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble, "Semi-supervised protein classification using cluster kernels," *Bioinformatics*, vol. 21, no. 15, pp. 3241–3247, 2005.

[14] E. Pacharawongsakda and T. Theeramunkong, "Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC," *NanoBioscience, IEEE Transactions on*, vol. 12, no. 4, pp. 311–320, Dec 2013.

[15] Q. Xu, D. H. Hu, H. Xue, W. Yu, and Q. Yang, "Semi-supervised protein subcellular localization," *BMC Bioinformatics*, vol. 10.

[16] J. K. Kim and S. Choi, "Probabilistic models for semisupervised discriminative motif discovery in DNA sequences," *IEEE/ACM Transactions on Computational Biology and Bioinfo.*, vol. 8, no. 5, pp. 1309–1317, 2011.

[17] L. Cerulo, C. Elkan, and M. Ceccarelli, "Learning gene regulatory networks from only positive and unlabeled data," *BMC Bioinformatics*, vol. 11, no. 1, pp. 228+, May 2010.

[18] Z. You, Z. Yin, K. Han, D.-S. Huang, and X. Zhou, "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network." *BMC Bioinformatics*, vol. 11, no. 1, p. 343, 2010.

[19] Y. Qi, 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, pp. i645–i652, 2010.

[20] K. Kundu, F. Costa, M. Huber, M. Reth, and R. Backofen, "Semi-supervised prediction of SH2-peptide interactions from imbalanced high-throughput data," *PLoS One*, vol. 8, no. 5, p. e62732, 2013.

[21] E. Kondratovich, I. I. Baskin, and A. Varnek, "Transductive support vector machines: Promising approach to model small and unbalanced datasets," *Molecular Informatics*, vol. 32, no. 3, pp. 261–266, 2013.

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

[23] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of the Eleventh Annual Conf. on Computational Learning Theory*, ser. COLT' 98. ACM, 1998, pp. 92–100.

[24] Q. Wei and R. L. Dunbrack Jr, "The role of balanced training and testing data sets for binary classifiers in bioinformatics," *PLoS One*, vol. 8, no. 7, p. e67863, 2013.

[25] A. Stanescu and D. Caragea, "Semi-supervised self-training approaches for imbalanced splice site datasets," in *Proc. of The Sixth Intl. Conf. on Bioinformatics and Computational Biology*, 2014, pp. 131–136.

[26] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch, "Accurate splice site prediction using support vector machines," *BMC Bioinformatics*, vol. 8, no. Suppl 10, pp. 1–16, 2007.

[27] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of The Thirty Third Annual Meeting on Assoc. for Computational Linguistics*, 1995, pp. 189–196.

[28] T. G. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*, ser. MCS '00. Springer-Verlag, 2000, pp. 1–15.

[29] V. Ng and C. Cardie, "Weakly supervised natural language learning without redundant views," in *Proceedings of The 2003 Conference of The North American Chapter of The Assoc. for Computational Linguistics on Human Language Technology*, vol. 1, 2003, pp. 94–101.

[30] G. Schweikert, C. Widmer, B. Schölkopf, and G. Rätsch, "An empirical analysis of domain adaptation algorithms for genomic sequence analysis," in *Advances in Neural Information Processing Systems 21*, vol. 8, 2008, pp. 1433–1440.

[31] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of The Twenty Third International Conference on Machine Learning*. ACM, 2006, pp. 233–240.