Semi-supervised self-training approaches in small and unbalanced datasets: Application to Xerostomia radiation side-effect

Abstract— **Supervised learning algorithms have been widely used as predictors and applied in a myriad of studies. The accuracy of the classification algorithms is strongly dependent on the existence of large and balanced training sets. The existence of a reduced number of labeled data can deeply affect the use of supervised approaches. In these cases, semi-supervised learning algorithms can be a way to circumvent the problem.**

In the present study, we apply several semi-supervised learning methodologies to a small clinical dataset with 222 (138 labeled and 84 unlabeled) head-and-neck cancer patients treated at the Portuguese Institute of Oncology of Coimbra (IPOCFG) with Intensity Modulated Radiation Therapy (IMRT). In order to predict the aptness for xerostomia induced by radiation treatments, we considered random forest classifiers. Xerostomia is one of the most frequent long term side-effects experienced by head-and-neck cancer patients undergoing radiation therapy, reducing drastically their quality-of-life. Therefore, being able to predict xerostomia at early stages of the treatment would make it possible to adjust the treatment plan in order to minimize or avoid this complication

The quality of the semi-supervised classification rule was validated by using different subsets of patients. Our experiments evidenced an improved performance of the classifier as the size of the training labeled dataset increased.

Keywords—**radiotherapy, xerostomia, semi-supervised learning, small databases, unbalanced datasets**

I. Introduction

In many research fields, the present challenge is not to obtain structured data but to develop machine learning algorithms capable of retrieving knowledge from the existing data. One of the most used applications is the development of classifiers. Classifiers can be trained using supervised learning (SL) algorithms, by considering a set of labeled samples and a set of explanatory attributes. Each sample is defined by a *p*-dimensional attribute vector and a class label, also known as response. The goal of the algorithms is the construction of a model (predictor or classifier), that allows to accurately predict the class labels of samples for which only the attribute vector is known [1]. In spite of SL being a very efficient technique, the production of accurate classifiers depends on the quality and quantity of labeled data. SL approaches require large labeled datasets in order to produce more useful and accurate classification rules [2]. When the working dataset is of reduced size, and there is a lack of labeled data, the application of supervised machine learning algorithms can thus be jeopardized. In many domains, obtaining labeled data is a particularly problematic task. One of the most appealing ways of circumventing the problem of the limited number of labeled samples is through automated semi-supervised learning (SSL) [2-5]. SSL uses both labeled and unlabeled data to construct a decision rule in order to improve a classifier trained only on the labeled pool. Classically, SSL uses large amounts of unlabeled data and small volumes of labeled data. Ideally, the training dataset should be sufficiently large and balanced in order to produce a classification model that outperforms a random decision rule [2]. To assure that the inclusion of unlabeled data will not worsen the performance of the classifier, it is crucial to consider rules that will define whether or not a given sample should be considered in the training process.

SSL methods can deal with the lack of labeled samples, but if there is a significant unbalance in the available dataset, this can lead to a biased learning. In fact, a non-uniform classes' distribution can lead to a partial learning, i.e., a model trained on an unbalanced dataset can tend to ignore the minority class, predicting samples as belonging to the majority one [4]. Therefore, in non-balanced class datasets, alternative solutions have been incorporated in both SL and SSL algorithms, either at data level, such as under and over sampling, or algorithm level, like cost-sensitive, active learning or even ensemble methods [2-8].

In the current paper, a medical application has been considered with the objective of predicting radiation-induced complications in the salivary glands for head-and-neck cancer patients treated with IMRT. The available dataset is both small and unbalanced.

Radiotherapy is one of the main treatments used against cancer, since cancer cells are less capable of repairing themselves than healthy cells if damaged by radiation. IMRT is one technique of radiation treatment that allows the achievement of a high degree of conformity between the area to be treated and the dose absorbed by healthy tissues [9]. The planning of a radiotherapy treatment is patient dependent, resulting in a trial and error procedure until a treatment complying as most as possible with the medical prescription is found. In spite of improvements gained with IMRT technique in head-and-neck cancer patients compared to old radiation therapy techniques, sparing of the salivary glands is still very challenging and the irradiation of such organs at risk can result in salivary dysfunction. Xerostomia, characterized by the feeling of dry mouth due to the lack of saliva, reducing drastically the quality-of-life of patients due to the difficulties in swallowing and in feeding, is one of the most frequent long term side-effects experienced by head-and-neck cancer patients undergoing radiation therapy [10]. Therefore, being able to predict xerosto-

mia prior to the radiation treatment can make it possible to optimize the treatment plan in order to minimize or avoid such complication. We have developed a xerostomia prediction model based on a labeled dataset of 138 patients with head-and-neck cancer treated at IPOCFG with IMRT, using random forest predictors. The SL model considers dosimetric information, namely, the planned mean dose in both parotids, and also specific patient features known prior to treatment age, gender and severity of xerostomia prior to radiation therapy treatment. These attributes revealed to be highly relevant predictors of xerostomia induced by radiation using random forests. One of the drawbacks of this approach is the fact that many unlabeled clinical cases cannot be considered. In the present study, we present a different approach by using SSL techniques that will allow the incorporation of more 84 unlabeled patients in the model construction. To the best of our knowledge, this is the first time that SSL algorithms are used to predict the risk for xerostomia induced by radiation treatments.

II. Material and methods

A. Dataset

This study considers a small and unbalanced clinical dataset with 222 head-and-neck cancer patients treated with IMRT. The patients' clinical data were exported from the electronic health information system RESPONSE [11] and include a number of patient features and medical registrations, such as patient and tumor characteristics, treatment details and patient response to radiation therapy registered during the follow-up medical consultations.

The classification of a side-effect at IPOCFG is made using RTOG/EORTC guidelines. A complication severity is ranked from 0 to 5, where 0 means no complication and 5 death from toxicity [12]. In the present study, we are not interested on the complication degree but only in predicting if a patient will develop or not xerostomia after radiation therapy. Therefore, only two severity classes were considered, namely "1" if the patient presented xerostomia and "0" otherwise. In the considered dataset, 84 samples are not labeled and from the 138 samples of the labeled pool, 86 correspond to patients with xerostomia, belonging to class "1", and 52 are complication-free, belonging to class "0".

The main purpose of the present work is to assess the improvements that can be obtained by incorporating unlabeled data in the training algorithm, compared with a SL classification approach. The features set considered in this study comprises dosimetric information, specifically, the planned mean dose in both parotids, age, gender and severity of xerostomia before irradiation.

B. Classification Model

We have considered the random forest prediction model to classify new patients according to the aptness for xerostomia 12 months after IMRT treatments. Random Forests works as an ensemble of decision tree classifiers [13], where leaves represent class labels and branches represent combinations of features leading to those class labels. The key of this procedure comprises the random selection of features to build a number of trees with locally-optimal decisions at each node. The split in each node is made according to the best feature among all possible features on the selected subspace. The class assigned to a new observation is the mode of the classes outputted by the individual trees.

R software was used, namely the "randomForest" R library [14]. In the predictive model, random forests are composed by 500 trees.

C. Semi-supervised Approaches

We have applied the self-training algorithm defined in [3], which starts by creating a prediction model trained on the labeled data. Then, the model is used to classify the unlabeled observations. The process is iterative, being the most confidently newly labeled samples added to the labeled dataset and the classifier re-trained, i.e., the topmost surely elements classified by the model in each step of the algorithm are then also used to self-train the model in the subsequent steps. The main goal of such procedure is to amplify the labeled dataset in order to produce a better-quality model. The different self-training approaches considered in the present study are briefly described below:

Self-Training with Unbalanced Dataset (STUD): In this algorithm, the original class distribution of the labeled dataset is maintained, in agreement with the approach developed by [3]. The newly labeled topmost confident instances are added to the labeled pool according to its positive-to-negative ratio. The performance of STUD algorithm is compared with the specific case of adding only the best classified sample (positive or negative) and also assuming an adding ratio 1:1, i.e., extracting from the unlabeled seed set the top most confident instances from each class, according to [4].

Self-Training with Over-Sampling (STOS): This algorithm starts by balancing the classes of the original labeled dataset. The instances of the majority class (the positive class in our dataset) are kept and the minority class (the negative one) is randomly over-sampled until an equal proportion between both classes is reached [3].

Self-Training with Under-Sampling (STUS): Similarly to STOS, STUS first balance the classes of the labeled seed set. In STUS, all instances of the minority class are kept and

the elements of the majority class are randomly picked without replacement until a balanced dataset is obtained [3, 4].

Self-Training with Under-Sampling Ensemble (STUSE): STSUE is a self-training algorithm with an ensemble approach. Several ensemble variants can be seen in [2-4, 6]. In general, multiple sets of initial training data are considered to train multiple classifiers, which work together as an ensemble to select confident elements from the unlabeled dataset to be included in the labeled pool. Indeed, many weak predictors self-trained on different subsamples of the labeled data can outperform the multi-view training [4]. Each classifier is trained on a balanced dataset containing all the minority instances and an equal number of majority elements randomly sampled without replacement. All training sets in the ensemble contain the same minority samples and different overlapping majority instances. A different classifier is trained with each of the balanced subsets, electing the two instances to be included into the labeled dataset for self-training in the next steps of the algorithm. The newly labeled instances with majority vote are selected for inclusion in the labeled datasets of each classifier and then the predictor models are re-trained. As the number of iterations increases, both the training sets and the ensemble models start to converge.

The first approach is not specifically designed to deal with unbalanced datasets, contrarily to the remaining ones. For the last three variants, the training labeled dataset is first balanced and thus, at each iteration of the algorithm, only two of the newly labeled elements are added to the labeled pool, the top most confident from each class. These approaches were designed to address the unbalanced data problem. Re-sampling helps to readjust the class distribution and thus the prediction model has an equal chance of learning the positive and negative classes.

Each classifier produces a value belonging to [0,1], corresponding to the probability of a patient belonging to a specific class. Classifying a patient will thus require this probability to be translated into a binary output: 1 or 0. This can be done by considering a threshold value α such that if the probability is greater than α, the assigned class should be “1”, and “0” otherwise. Each α represents a decision boundary in the feature space. The most used threshold value is the value 0.5. However, tightening the decision boundary can lead to more certainty and accurate classifications and thus enhance the performance and quality of the classifier. In the present study, we have chosen to classify an unlabeled instance as positive if the probability value yielded by the model is equal or greater than 0.8; in case of such value is equal or smaller than 0.2 the sample is classified as negative. All other instances resulting in model values outside this range were discarded from the analysis in order not to affect the quality of the results. This means that all unlabeled samples used in the training of the classifier are being classified with a high level of certainty and thus will contribute to a more consistent predictor.

D. Evaluation Metrics

The usual rule of thumb to create a best sample and improve the performance of the classifier when dealing with small datasets is producing a "bootstrap sample", which is a sample higher or equal in size to the original dataset but generated by random selection with replacement [15]. In the present study, we run the classifier without and with bootstrap samples, generating sets with 500 elements in the latter case. Regarding the composition of the classes, we considered two bootstrapping situations: random bootstrap sampling and balanced bootstrap samples. Self-training is considered by many authors a bootstrapping method. However, in the present work this technique is not used with this purpose.

The most commonly recommended and used performance measures to judge the discriminative ability of a model when handling binary outcomes are the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) [16]. Nevertheless, many authors consider the Area Under the Precision-Recall Curve (AUPRC) a more appropriate assessment metric when undertaking problems with unbalanced datasets [17]. Therefore, in the current study, we evaluate the performance of the classifier by using both, AUC and AUPRC.

In order to assess the suitability of the prediction model, we applied a cross-validation procedure [18]. Cross-validation consists in splitting the original dataset into complementary subsets: a training set, used to perform the analysis, and a testing or validation set, used to confirm the results. In this study, 20 labeled samples are used to validate the analyses, being the remaining ones used to build the prediction model. For each SSL approach, the algorithm was run 100 times, calculating at the end the average and the standard deviation of the obtained performance measures.

Table 1 Small and unbalanced dataset.

	SSL (Mean±SD)			SL (Mean±SD)	
	AUC	*AUPRC*	*New Labels*	*AUC*	*AUPRC*

STUD					
Proportion	**0.71±0.11**	**0.82±0.09**	11±9		
Best	0.69±0.12	0.80±0.11	36±8	0.69±0.12	0.80±0.11
1:1	**0.72±0.11**	**0.82±0.09**	8±7		
STOS	0.71±0.10	**0.74±0.09**	19±7	0.71±0.10	0.73±0.11
STUS	0.66±0.11	0.70±0.11	24±8	0.69±0.12	0.72±0.12
STUSE	0.69±0.10	**0.73±0.10**	21±6	0.70±0.11	0.71±0.11

III. Results and discussion

The performances of SSL methodologies were compared with the correspondent SL ones in order to understand the benefits of increasing the training labeled dataset by using unlabeled samples, assuming the same methodological conditions. Table 1 displays the results obtained when applying random forests to the original small and unbalanced dataset. As it can be easily observed, SSL approaches outperform SL methodologies in almost all situations. Moreover, in all cases, AUPRC measure revealed better than AUC. Such evidence is most notorious in the cases that deal with unbalanced training sets, which is the case of STUD approach. The disproportion between classes is not highly significant as well as the number of unlabeled data is not too large; so, the results obtaining when applying the STUD approach adding the new labeled samples to the labeled dataset in proportion with the unbalance ratio or according to the 1:1 ratio are similar. The performance of the model is lightly improved by using SSL in both cases with a few number of new quality labeled samples added to the labeled seed set. When considering the inclusion of only the best new classification, no improvement is achieved, even adding more than a few samples. It could be related with the fact that a model trained on an unbalanced dataset tends to better classify the samples of the majority class and the labeled dataset may be getting increasingly unbalanced. When applying machine learning classification techniques specifically designed to deal with non-uniform class distributions, only the AUPRC measure is able of producing enhanced results by SSL compared with SL. From the three used variants that first readjust the class distribution, only the STUS approach results in a worse-quality performance by SSL. Undersampling the majority class can result in an even smaller and poorly diversified dataset, skewing the learning of the model. However, considering an ensemble approach of overlapped undersamples of the majority class, as adopted by STUSE algorithm, the SSL analyses are improved, outperforming the SL ones. This approach considers several weak classifiers that are trained on different sample sets. A multi-view training allows a more diversified analysis and thus more precise class learning. Several classifiers working together for the same purpose can execute better tasks than individual predictors. On the other hand, oversampling the minority class can result in a larger training set with a greater sample variance, which leads to an easier learning, resulting thus in a better performance when applying the STOS methodology by SSL.

Table 2 and 3 illustrate the performance of SSL and SL algorithms when considering the generation of bootstrapping samples to train the prediction models. Only the STUD approach was run, since a bootstrapping technique does not enrich the training data of the remaining methodologies. Assuming random bootstrapping samples, no STUD variant is improved by using SSL instead of SL. Random bootstrapping may increase the disparity between the compositions of the classes introducing bias in the learning of the model. The real disparity of the classes is not very large, but it may be accentuated by random bootstrapping, negatively affecting the performance of the model. In contrast, when generating balanced bootstrap samples, all STUD alternatives result in a better performance by SSL, which is more evidenced with the AUPRC measure. The performance of the classifier is lightly improved by the inclusion of some new labeled instances.

In general, the performance of a classifier can be improved by increasing the size of the training labeled set by SSL approaches, either dealing with the original small and unbalanced labeled set, adding the instances according to the class proportion or the 1:1 ratio, or adjusting the initial class distribution by using re-sampling techniques or generating balanced bootstrapping samples.

Table 2 Random bootstrapping samples.

	SSL (Mean±SD)			SL (Mean±SD)	
	AUC	*AUPRC*	*New Labels*	*AUC*	*AUPRC*
STUD					
Proportion	0.71±0.11	0.81±0.09	25±9	0.73±0.11	0.83±0.11

Best	0.68±0.11	0.79±0.10	62±8	
1:1	0.72±0.13	0.82±0.10	18±8	

Table 3 Balanced bootstrapping samples.

	SSL (Mean±SD)			SL (Mean±SD)	
	AUC	*AUPRC*	*New Labels*	*AUC*	*AUPRC*
STUD					
Proportion	**0.71±0.11**	**0.82±0.09**	32±6		
Best	**0.71±0.10**	**0.82±0.09**	43±8	0.70±0.12	0.80±0.11
1:1	**0.71±0.12**	**0.82±0.11**	30±8		

IV. Conclusions

In this paper a medical-case study was considered where it was necessary to address the problem of having only a small sized dataset with an unbalanced class distribution. As far as the authors know, this is the first time that SSL algorithms are used in the context of predicting the aptness for xerostomia after radiation therapy. Our empirical results revealed a successful utilization of the unlabeled data by SSL approaches. In small and unbalanced datasets the performance of the SSL classifier is slightly improved when compared with the use of SL methodologies. The use of labeled data is enriched by the unlabeled one. The smaller the dataset, the greater the difficulty in describing satisfactorily the data patterns. Therefore, by increasing the size of the dataset, we would be able of highly improving the performance of the classification model. Furthermore, classifiers are negatively affected when learning from datasets with non-uniform distributions. The more skewed the distribution of the classes, the more affected the performance of the classifier is. Our experiments show that classifiers can have better performances on unbalanced datasets if they are self-trained according to the original data set unbalance or 1:1 ratios or on balanced datasets, either by resampling techniques or by balanced bootstrapping samples.

In future work we intend to develop a methodology that generates synthetic data in order to increase the amount of labeled data as well as balancing the classes' distribution. Moreover, we are also interested in applying co-training algorithms and possibly explore cost-sensitive techniques.

Acknowledgment

This work was supported by FEDER, COMPETE, iCIS (CENTRO-07-ST24-FEDER-002003), and also PTDC/EIA-CCO/121450/2010, PEst-OE/EEI/UI308/2014.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Culp M, Johnson K, Michailidis G (2006). ada: An R Package for Stochastic Boosting. J Stat Soft. 17:
2. Stanescu A, Caragea D (2014). Ensemble-based semi-supervised learning approaches for imbalanced splice site datasets. 432-437.
3. Stanescu A, Caragea D (2014). Semi-supervised self-training approaches for imbalanced splice site datasets. Proceedings of the 6th International Conference on Bioinformatics and Computational 131-136.
4. Li S, Wang Z, Zhou G et al. (2011). Semi-supervised learning for imbalanced sentiment classification. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. 3:1826-1831
5. Tangirala K, Caragea D (2011). Semi-supervised Learning of Alternatively Spliced Exons Using Co-training. 243-246.
6. Liu X-Y, Wu J, Zhou Z-H (2009). Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 39:539-550.
7. Li S, Ju S, Zhou G et al. (2012). Active learning for imbalanced sentiment classification. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 139-148.
8. Ling CX, Sheng VS, Cost-Sensitive Learning and the Class Imbalance Problem, in Encyclopedia of Machine Learning, C. Sammut (Ed.). 2008, Springer.
9. Lee NY, Terezakis SA (2008). Intensity-modulated radiation therapy. Journal of Surgical Oncology. 97:691-6.
10. Messmer MB, Thomsen A, Kirste S et al. (2011). Xerostomia after radiotherapy in the head & neck area: long-term observations. Radiother Oncol. 98:48-50.
11. Ferreira BC, Khouri L, Lopes MC et al. (2015). RESPONSE, an Electronic Health Patient Information Software for Radiation Therapy. Proceedings of the 6th Europ Conf of the Int Fed for Med and Biol Engin. 45:691-4.
12. Cox JD, Stetz J, Pajak TF (1995). Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys. 31:1341-6.

13. Breiman LEO (2001). Random Forests. Mach Learn. 45:5-32.
14. Liaw A, Wiener M (2014). Package 'randomForest'.
15. Efron B, Tibshirani R. An introduction to the bootstrap. 1st CRCPress reprint (Eds.) Boca Raton: Chapman & Hall/CRC 1998 237-81.
16. Fawcett T. ROC graphs: notes and practical considerations for data mining researchers. Technical report hpl-2003-4 (Eds.) HP Laboratories 2003 Palo Alto, CA, USA.
17. Davis J, Goadrich M (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning. 233 -240.
18. Molinaro AM, Simon R, Pfeiffer RM (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics. 21:3301-7.