Editorial version available at: https://doi.org/10.1016/j.eswa.2020.113569

Semi-supervised regression trees with application to QSAR modelling

Jurica Levatić[a,b,*], Michelangelo Ceci[c], Tomaž Stepišnik[a,b], Sašo Džeroski[a,b], Dragi Kocev[a,b]

[a] *Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia*
[b] *Jožef Stefan International Postgraduate School, Ljubljana, Slovenia*
[c] *Department of Computer Science, University of Bari Aldo Moro, Bari, Italy*

Abstract

Despite the ease of collecting more and more data about various phenomena, obtaining labeled data needed for learning models with high predictive performance remains a very difficult and expensive task. To leverage the information from the small amounts of labeled data, we need to also exploit the information coming from unlabeled data. This can be achieved by employing semi-supervised machine learning methods.

In this paper, we propose a novel semi-supervised method that learns interpretable regression trees. It is based on the predictive clustering trees paradigm that extends regression trees towards structured output prediction. We also propose to learn ensembles of semi-supervised regression trees.

The method we propose is particularly suited for the chemoinformatics task of quantitative structure-activity relationship (QSAR) modeling, which is the main application context considered in this paper. Specifically, we evaluate the proposed method on 4 QSAR modelling datasets and illustrate its use in a case study. Additionally, we also evaluate our approach on 8 benchmark datasets not related to the QSAR modeling problem.

The evaluation reveals the following: semi-supervised trees and ensembles

*Corresponding author
Email addresses: Jurica.Levatic@ijs.si (Jurica Levatić), michelangelo.ceci@uniba.it (Michelangelo Ceci), Tomaz.Stepisnik@ijs.si (Tomaž Stepišnik), Saso.Dzeroski@ijs.si (Sašo Džeroski), Dragi.Kocev@ijs.si (Dragi Kocev)

have better predictive performance than their supervised counterparts (especially when the number of labeled examples is very small); different datasets and different amounts of labeled data require different amounts of unlabeled data to be included in the learning process; and the learned semi-supervised regression trees can be used for better understanding the problem at hand and the way predictions are being made.

Keywords: Semi-supervised learning, Regression, Decision trees, Random forests, QSAR

1. Introduction

Semi-supervised learning (SSL) (Chapelle et al., 2006) aims to leverage machine learning algorithms by exploiting both labeled and unlabeled data. The motivation to include unlabeled data in the learning process stems from the fact that labeled data are hard to obtain in many applications of machine learning, while unlabeled data are easily available in large quantities. For example, in chemistry, determination of the biological activity of compounds requires expensive and time-consuming experiments, while a huge amount of unlabeled compounds is freely available through public databases (such as ChEMBL[1] or PubChem[2]), c.f., Figure 1. Furthermore, in ecology the presence and abundance of specific species at a given site has to be manually determined by experts, while descriptive attributes for the site, such as temperature or humidity, are easy to obtain. The lack of labeled data in domains such as the above-mentioned can negatively affect the predictive performance of supervised machine learning algorithms, highlighting the need for semi-supervised learning.

Orthogonal to the ability to exploit unlabeled data, another feature of machine learning algorithms that is often necessary is the interpretability of the models learned. This feature is particularly important given that many (or

[1] `https://www.ebi.ac.uk/chembl/`

[2] `https://pubchem.ncbi.nlm.nih.gov/`

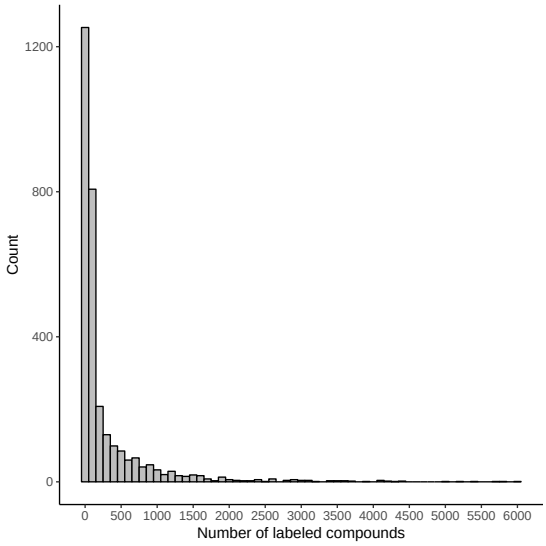


Figure 1: Histogram of dataset sizes (in terms of number of labeled compounds) for 3047 biological targets extracted from the ChEMBL database. For a vast majority of targets, less than 100 compounds are labeled. These QSAR datasets are available for download from OpenML (Vanschoren et al., 2014).

probably most) of the developed machine learning tools do not natively generate interpretable models, whereas machine learning techniques able to explain the decisions of the models are increasingly important in many domains (Escalante et al., 2017). Practitioners of machine learning are frequently not only interested in the predictive performance of the models, but also apply them to extract knowledge from data and generate novel plausible hypotheses. The prominent machine learning methods bearing this property are decision trees (Breiman et al., 1984) and rules (Fürnkranz et al., 2012).

Semi-supervised learning has emerged several decades ago as a sub-field of machine learning (Chapelle et al., 2006). Since then, a plethora of semi-supervised learning algorithms have been proposed in the literature. However, the majority of the development has been focused on the classification task, while the regression task has received much less attention (Zhu, 2008). Furthermore, we are not aware of machine learning methods for regression that

can learn both semi-supervised and interpretable models - apart from the self-training wrapper approach proposed by Yarowsky (1995).

In this paper, we propose an algorithm for semi-supervised learning of regression trees. The algorithm employs unlabeled data directly in the tree construction phase, inheriting the properties of regression trees, i.e., interpretability and low computational complexity. Additionally, the algorithm adapts to the data at hand by controlling the influence of unlabeled (relative to labeled) data, lowering the risk of unlabeled data degrading the predictive performance. Furthermore, semi-supervised ensembles can easily be constructed by using the proposed semi-supervised regression trees as base learners.

The application domain considered in this study is the chemoinformatics task of quantitative structure-activity relationship (QSAR) modeling. In QSAR modeling, the goal is to relate a description of a molecule's structure and properties to its biological activities. Conceived as an extension of physical organic chemistry in 1960s, QSAR modeling has since grown to become a standard requisite in the drug development process. Stimulated by the continous growth of chemical data and databases, QSAR modeling has evolved from the analysis of small series of similar compounds using simple regression methods, to the analysis of large datasets of diverse molecular structures using a wide variety of statistical and machine learning techniques (Cherkasov et al., 2014).

Machine learning techniques have been able to improve over the simple linear regression methods initially used in deriving QSAR models. However, existing approaches do not consider (at least not simultaneously) two important challenges in the domain of QSAR modelling:

- Labeled datasets are typically very small, because of the complex and expensive annotation process, often containing only few tens of compounds. For example, the largest QSAR study to date considered 2764 datasets, where median dataset size is 73 (Olier et al., 2018). On the contrary, unlabeled data are abundant in public databases.

- QSAR modeling has shifted away from simple and interpretable models

towards more complex multiparametric approaches, somewhat trading interpretability for better predictive ability. However, interpretability is still desired, in order to facilitate the practical acceptance of QSAR solutions by domain experts (Cherkasov et al., 2014).

The semi-supervised learning algorithm of regression trees that we propose in this paper addresses both challenges. The empirical evaluation of the proposed approach demonstrates its validity and its relevance for QSAR modelling, as well as for other application domains. In the remainder of this paper, we describe the proposed method (Section 2), lay out the experimental questions and experimental design (Section 3), present and discuss the results of the experimental evaluation (Section 4), show a case study on predicting inhibitors of farnesyltransferase (Section 5), discuss work related to this study (Section 6), and finally conclude the paper with a summary of the main findings (Section 7).

2. Semi-supervised learning of regression trees

In classical regression, given a set of observed data $\{x, y\} \in X \times Y$, where X denotes the feature space spanned by m independent (or predictor) variables x_i (both numerical and categorical), the goal is to predict the target variable Y that is continuous.

A regression tree approximates a function $y = g(x)$ by means of a piece-wise constant function. The standard way to construct this function is top-down induction, that is, recursive partitioning of the training set while moving down the tree (Breiman et al., 1984). The output model is a tree, where each internal node represents a partitioning of the example space and each leaf represents a subdomain (and its associated constant function) of the piece-wise constant function. In such top-down construction of a regression tree, one of the main problems is choosing the best partition (or split) of a region of the example space. For this purpose, several evaluation functions have been proposed in the literature, mostly based on the mean square error of the response variable (Malerba et al., 2004): The main idea is to choose the partition that minimizes

the mean square error (computed on the training set) of the resulting piece-wise constant function.

A more general formulation of this solution is that of choosing the best partition of a region of the example space by considering, according to the semi-supervised learning setting, not only the labeled examples, but also unlabeled examples. This, however, requires a different evaluation function that not only depends on the response variable (whose values are not available for unlabeled examples), but also depends on the values of predictor variables, whose values are available for both labeled and unlabeled examples.

Following this generalization, in this work, we propose an algorithm for learning semi-supervised regression trees. Their construction is still based on the top-down induction algorithm. However, there is a major difference with respect to standard regression tree induction algorithm: the evaluation function (h) evaluates the impurity (imp) of candidate splits by taking into account both labeled and unlabeled examples.

More formally, the best split (partition) is chosen such that the impurity reduction is maximized, i.e., the difference between the impurity of the parent node and the sum of impurities of child nodes:

$$h = imp(E) - \left(\frac{|E_y|}{|E|} imp(E_y) + \frac{|E_n|}{|E|} imp(E_n) \right) \tag{1}$$

where E, E_y, E_n are the sets of examples in the parent, left and right child nodes, respectively.

In *supervised* regression trees, the impurity measure corresponds to the variance of target variable values belonging to the examples in a given node:

$$imp(E) = Var(E) = \frac{\sum_{i=1}^{N} (y_i)^2 - \frac{1}{N} \cdot \left(\sum_{i=1}^{N} y_i \right)^2}{N}, \tag{2}$$

where y_i is the value of the target (response) attribute of the i^{th} example, and $N = |E|$ is the number of examples.

Classical supervised regression tree learning algorithms evaluate the splits considering only the target variable, i.e., the examples in the resulting nodes

(partitions) are homogeneous with respect to the target variable. In this work, we propose semi-supervised regression trees that produce nodes homogeneous both with respect to the target variable and the descriptive attributes.Our theoretical rationale follows the *semi-supervised smoothness assumption*, which states that if two examples e_i and e_j in a high-density region are close (with respect to descriptive space), then also their target values y_i and y_j should be close (Chapelle et al., 2006). By including (a large quantity of) unlabeled examples into the learning process, the semi-supervised regression trees can produce splits that group together examples that are close both in the descriptive and the target space. Then, if the semi-supervised smoothness assumption holds, more accurate predictions can be achieved. It is noteworthy that, in order to benefit from unlabeled data (that by definition does not carry any information about the target variable), semi-supervised methods are bound to assumptions about the distribution of the unlabeled data with respect to the target variables (Chapelle et al., 2006).

The proposed semi-supervised learning of regression trees utilizes the same principle of impurity reduction maximization as in Eq. 1. The above-described extension to a semi-supervised setting is achieved by modifying the impurity function so that it accommodates both labeled E_l and unlabeled E_u examples. The impurity of a set of examples $E_{l+u} = E_l \cup E_u$ in semi-supervised trees is then computed as a weighted sum of impurities over the target variable and the independent attributes:

$$imp_{SSL}(E_{l+u}) = w \cdot \frac{imp^{Y}_{SSL}(E_l)}{imp^{Y}_{SSL}(E^{train}_l)} + \frac{1-w}{D} \cdot \sum_{i=1}^{D} \frac{imp^{X_i}_{SSL}(E_{l+u})}{imp^{X_i}_{SSL}(E^{train}_{l+u})}, \qquad (3)$$

where Y is the target variable, D is the number of independent attributes, and X_i is the i^{th} independent attribute. To ensure commensurate contributions of the target and the descriptive attributes, normalization with the respective impurities over the entire training set (E^{train}) is performed.

The weight parameter $w \in [0, 1]$ controls the relative contribution of the target variable and the descriptive attributes to the overall impurity. It is noteworthy that the impurity over the target variable (imp^{Y}_{SSL}) is computed using only

the labeled examples (E_l), whereas the impurity over the descriptive attributes ($imp_{SSL}^{X_i}$) is calculated by using all the examples (i.e. labeled and unlabeled examples, E_{l+u}). Therefore, the w parameter controls the amount of supervision employed during the learning of semi-supervised trees, where increasing the value of w corresponds to more supervision and decreasing it to less supervision, enabling the learning of trees to range from perfectly supervised (i.e., $w = 1$) to perfectly unsupervised (i.e., $w = 0$).

The impurity over the target variable Y is calculated by using Eq. 2, while the impurity of the descriptive attributes depends on the type of the attribute:

$$imp_{SSL}^{X_i}(E_{l+u}) = \begin{cases} Var^{X_i}(E_{l+u}), \text{ if } X_i \text{ is numeric} \\ Gini^{X_i}(E_{l+u}), \text{ if } X_i \text{ is nominal,} \end{cases} \tag{4}$$

where $Var^{X_i}(E_{l+u})$ is the variance of a set of examples E_{l+u} for the attribute X_i calculated as defined in Eq. 2 and $Gini^{X_i}(E_{l+u})$ is the Gini score of a set of examples E_{l+u} for the attribute X_i calculated as follows:

$$Gini(E_{l+u}) = 1 - \sum_{i=1}^{C} p_i^2, \tag{5}$$

where C is the number of categories of the attribute X_i, and p_i is the apriori probability of the category c_i (i.e., the relative frequency of examples in the set E_{l+u} belonging to the category c_i).

Having defined semi-supervised regression trees, we can easily extend them to semi-supervised regression tree ensembles, by simply using the semi-supervised trees as base predictive models of an ensemble. In this work, we extend the random forest algorithm (Breiman, 2001) to a semi-supervised setting in this way. More specifically, we construct ensembles of semi-supervised predictive models by creating bootstrap replicates of the training set and using each replicate to construct a semi-supervised predictive model. Additionally, according to the random forest approach (Breiman, 2001), the set of descriptive attributes considered for split selection in each node is a freshly randomized selection of the attributes. The prediction of an ensemble for a new instance is then obtained

by combining (averaging) the predictions of all the base predictive models from the ensemble. The semi-supervised regression trees and semi-supervised random forests we propose in this work are based on the predictive clustering framework, namely predictive clustering trees (PCTs) (Blockeel et al., 1998) and ensembles thereof (Kocev et al., 2013). Both semi-supervised trees and random forests thereof are implemented in the CLUS system (Blockeel & Struyf, 2002) and are available at `http://kt.ijs.si/jurica_levatic/`.

We next discuss the computational complexity of the proposed method. We start by discussing the computational complexity of learning a supervised PCT. Learning a supervised PCT requires the following steps which contribute to the computational complexity as follows: sorting the values of D descriptive attributes ($\mathcal{O}(DN \log N)$), calculating the best split ($\mathcal{O}(DN)$), and applying the split to the N (labeled) training examples ($O(N)$). Assuming that the depth of the tree is $\mathcal{O}(\log N)$ (Witten & Frank, 2005), the computational complexity of constructing a single (supervised) PCT is $\mathcal{O}(DN \log^2 N) + \mathcal{O}(DN \log N) + \mathcal{O}(N \log N)$.

Learning a semi-supervised PCT involves both labeled and unlabeled examples (i.e., $N = N_l + N_u$, instead of $N = N_l$). Also both the target variable and the D independent attributes are used when the splits are evaluated, thus the complexity of the evaluation of a single split is $\mathcal{O}((1 + D)DN)$. This gives the total computational complexity of learning a single semi-supervised PCT of $\mathcal{O}(DN \log^2 N)) + \mathcal{O}((1 + D)DN \log N) + \mathcal{O}(N \log N)$.

The worst case computational complexity of learning semi-supervised random forests is $k(\mathcal{O}(D'N' \log^2 N') + \mathcal{O}((1 + D)D'N' \log N'))$, where N' is the size of the bootstrap samples. D' is the size of the attribute subsets at each tree node, and k is the number of base models in the ensemble.

3. Experimental design

3.1. Data description

We evaluate the predictive performance of our approach using four QSAR datasets from the OpenML repository (Vanschoren et al., 2014). The tasks at hand are to relate molecular properties to biological activities: the Neurokinin 1 receptor, the Glycogen synthase kinase-3 alpha, the Rho-associated protein kinase 2 and the Human immunodeficiency virus type 1 protease, respectively. Moreover, we use 8 additional benchmark datasets that differ in the domains they represent, the number of attributes and the number of examples. These datasets are obtained from the Keel repository (Alcalá et al., 2010) and from the repository of Luís Torgo (Torgo, 2016). We have selected all the 12 regression datasets to have more than a thousand of examples, so that we can use an evaluation scenario relevant for SSL where large amounts of unlabeled data are available.

3.2. Experimental setting

In this work, we propose semi-supervised regression trees (SSL-PCT) and semi-supervised random forest tree ensembles (SSL-RF). We compare these methods to their supervised counterparts, i.e., regression trees (CLUS-PCTs), and random forests (CLUS-RF), respectively. These are the most reasonable baselines, as the purpose of the comparison is to evaluate the contribution provided by unlabeled data to the overall predictive capabilities in an experimental setting that guarantees a fair comparison.

In the experiments, both supervised and semi-supervised trees are pruned with the M5P pruning procedure (Quinlan, 1993). For each variant of the ensemble approaches (i.e., CLUS-RF and SSL-RF), we build random forests consisting of 100 unpruned trees. When building these trees, the number of randomly considered attributes, at each internal node, is set to $\lfloor \log_2(D) + 1 \rfloor$, where D is the number of independent attributes (Breiman, 2001).

In order to explore the influence of the amount of labeled data on the predictive capabilities of the semi-supervised methods, we perform experiments by

Table 1: Characteristics of the datasets. N: number of instances, D/C: number of descriptive attributes (nominal/continuous).

Dataset name and source	Domain	N	D/C
Neurokinin 1 receptor (NK1) (Vanschoren et al., 2014)	QSAR	2446	1024/0
Glycogen synthase kinase-3 alpha (GSK3A) (Vanschoren et al., 2014)	QSAR	1211	1024/0
Rho-associated protein kinase 2 (ROCK2) (Vanschoren et al., 2014)	QSAR	1521	1024/0
Human immunodeficiency virus type 1 protease (HIV-1) (Vanschoren et al., 2014)	QSAR	4442	1024/0
2dplanes (Torgo, 2016)	Artificial	40768	0/10
Abalone (Torgo, 2016)	Biology	4177	1/7
Elevators (Torgo, 2016)	Optimal control	16559	0/18
Kinematics (Torgo, 2016)	Robotics	8192	0/8
Laser (Alcalá et al., 2010)	Optics	993	0/4
Plastic (Alcalá et al., 2010)	Plastic strength	1650	0/2
Pole (Torgo, 2016)	Telecommunication	5000	0/48
Stock (Torgo, 2016)	Economy	950	0/10

varying the absolute number of labeled examples in the set {25, 50, 100, 200, 350, 500}. The labeled examples used for the training phase (both for supervised and semi-supervised learning) are randomly sampled from the whole dataset. The remaining examples are then used both as unlabeled examples and as testing set. This evaluation approach is coherent with the transductive learning setting (Malerba et al., 2009) where the learning algorithms cannot see the labels of the examples that are considered "unlabeled" but, during evaluation, the evaluation measures are computed on the true labels. In order to guarantee a fair comparison between semi-supervised and supervised algorithms, the supervised models are learned exclusively on the labeled part of the data and their performance is evaluated on the same test data. Moreover, in order to guarantee valid results and conclusions, the whole evaluation procedure is executed 10 times. All the results reported in this paper refer to the average of the evaluation measures obtained with these 10 runs.

In the experiments, the values of the parameter w are automatically set according to a 3-fold cross-validation on both labeled and unlabeled examples of the training set. Specifically, for each run, the algorithm identifies the value of w, taken from the set $\{0, 0.1, 0.2, \cdots 1\}$, which optimizes the considered evaluation measure.

The predictive performances of the algorithms are evaluated in terms of the relative root-mean-square-error (RRMSE). In addition, we also perform a statistical evaluation of the results in order to statistically assess the differences among the considered and proposed methods. At this purpose, we use the Wilcoxon paired signed rank test (Wilcoxon, 1945) by comparing the average RRMSE of two methods over the considered datasets. In all the statistical tests reported in the following, the selected significance level is 0.05.

4. Results and discussion

In this section, we present the empirical results obtained. We first confirm the ability of the proposed semi-supervised methods to take advantage of unlabeled data by comparing them to their supervised counterparts in terms of predictive performance. We then analyze and discuss the characteristics of the proposed methods from the aspect of practical usability: sensitivity to parameters and interpretability of the trees.

4.1. Predictive performance

Figure 2 depicts how the predictive performance (RRMSE) of semi-supervised (SSL-PCT and SSL-RF) and supervised methods (CLUS-PCT and CLUS-RF) changes with increasing the amount of labeled data on the 12 regression datasets.

We can observe that SSL-PCTs achieve lower predictive error than CLUS-PCTs on several datasets: 2dplanes, Elevators, Kinematics, Laser and Pole. Furthermore, it seems that SSL-PCTs are especially effective on the QSAR datasets (NK1, GSK3A, ROCK2 and HIV-1) when the amount of labeled data

is rather limited (i.e., from 25 to 100 of labeled examples). This case corresponds to the size of the datasets typically used in QSAR studies.

Semi-supervised random forests achieve consistent improvement over CLUS-RF on the majority of datasets. Furthermore, it seems that SSL-RF can improve over CLUS-RF even if SSL-PCT does not improve over CLUS-PCT and vice versa (e.g., for Plastic and Stock datasets). In other words, the improvement provided by SSL is orthogonal to the improvement provided by ensembles (random forests). In particular, the domain of QSAR modeling seems to be suitable for SSL-RF, since SSL-RF improves over CLUS-RF for almost all different amounts of labeled data on all four QSAR datasets considered.

Table 2 presents a statistical analysis of the predictive performance estimates. We can observe that semi-supervised regression trees are the most effective for small amounts of labeled data: They achieve statistically significantly better performance than supervised PCTs up to 100 labeled examples. A similar observation can be made for semi-supervised random forests: They achieve statistically significantly better performance than supervised random forests for amounts of labeled data ranging from 25 to 350 examples, while statistical significance of improvement is not achieved for 500 labeled examples.

4.2. Controlling the influence of unlabeled data with the w parameter

As mentioned before, the w parameter controls the influence of the amount of information coming from unlabeled examples. This aspect enables semi-

Table 2: p-values obtained with the Wilcoxon signed-rank test on the RRMSE performance values of both supervised and semi-supervised algorithms. In bold, we indicate significant p-values (< 0.05) for which there is a statistically significant difference between the compared methods. In all of the comparisons, the semi-supervised algorithm outperformed its supervised counterpart in terms of average RRMSE and sums of ranks.

Methods			Number of labeled examples					
			25	50	100	200	350	500
CLUS-PCT	vs.	SSL-PCT	**0.011**	**0.010**	**0.004**	0.367	0.480	0.583
CLUS-RF	vs.	SSL-RF	**0.008**	**0.065**	**0.008**	**0.023**	**0.034**	0.126

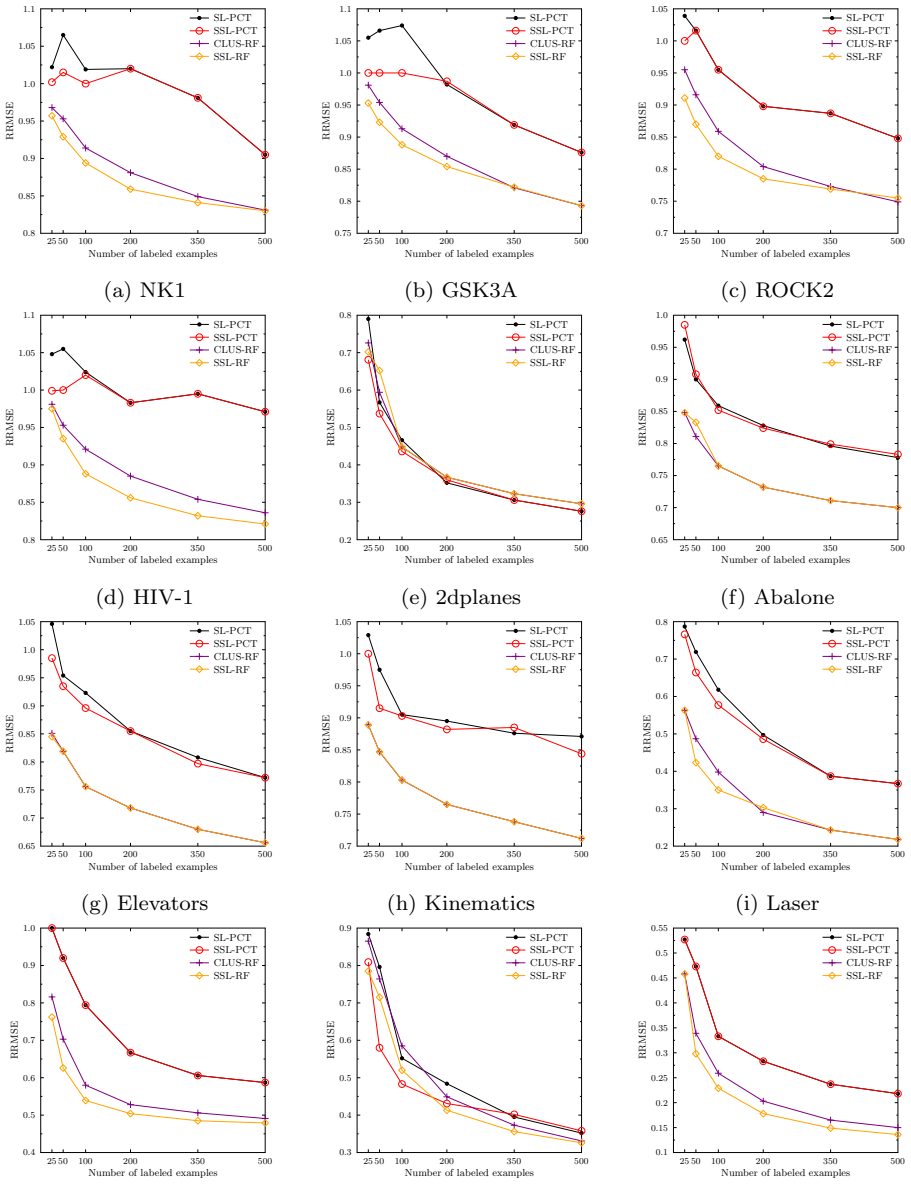


Figure 2: RRMSE of the supervised and semi-supervised methods on the regression datasets.

supervised trees to adapt to the data at hand, which is important since, as several studies have demonstrated, unlabeled examples are sometimes not beneficial for the prediction accuracy of semi-supervised algorithms (Cozman et al., 2002; Nigam et al., 2000; Guo et al., 2010; Zhou & Li, 2007).

Figure 3 presents the values of w, automatically determined by the internal cross-validation procedure (see Section 3). We can observe that the determined values of w vary among the datasets and even for the same dataset, with different number of labeled examples used. Therefore, a general recommendation for setting the value of w is difficult to provide, rather, the w parameter must be optimized for each dataset separately. As noted before, SSL-PCTs are most successful up to 100 labeled examples (Section 4.1 and Fig. 2). This is also reflected in the values of w: as the amount of labeled data increases, $w = 1$ is more frequently chosen as an optimal value for SSL-PCTs (Fig. 3, top panel). Note that $w = 1$ means that the algorithm disregards unlabeled examples and essentially performs supervised learning, likely preventing unlabeled examples to hurt the predictive performance. Also, some datasets do not benefit from semi-supervised learning (with the proposed algorithms), such as Stock for SSL-PCTs or Kinematics for SSL-RF, where $w = 1$ is used for all the cases, independently of the amount of labeled data.

Interestingly, for datasets where SSL-RF improves over CLUS-RF, $w = 0$ is almost always chosen as an optimal value for SSL-RF (Fig. 3-Random Forests, bottom panel). In these cases, the models are induced with unsupervised learning: the examples are clustered considering only the descriptive attributes, while the target variables are employed only to assign labels in the leaf nodes of the trees. This observation may suggest that the data in these datasets comply with the semi-supervised smoothness assumption (see Section 2) – the examples that are close in the descriptive space also have similar target values. This is particularly true for QSAR datasets, suggesting that SSL is particularly helpful in this specific application domain.

To further emphasize the importance of controlling the influence of unlabeled data, we compare the predictive performance of semi-supervised models induced

with the w chosen by internal cross-validation to the models where $w = 0.5$, that is, when an equal weight is given to unlabeled and labeled data (Figures 4). We can observe that the former models, which adapt to the dataset at hand, achieve better predictive performance in almost all of the cases.

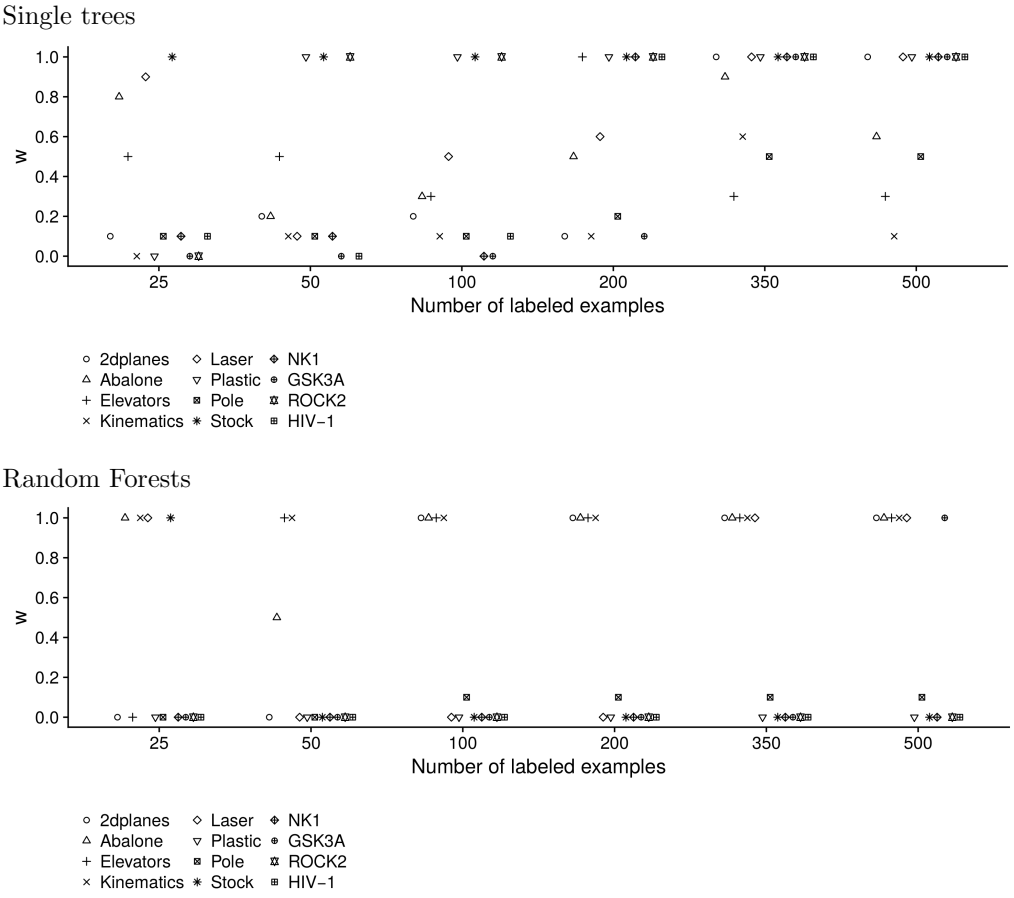


Figure 3: An illustration of the w parameter values used in the experiments (as chosen by cross-validation) for different datasets and various amounts of labeled data used.

4.3. Interpretability of the learned trees

Interpretability is a desired property of predictive models in many applications of machine learning. The semi-supervised PCTs method learns predictive

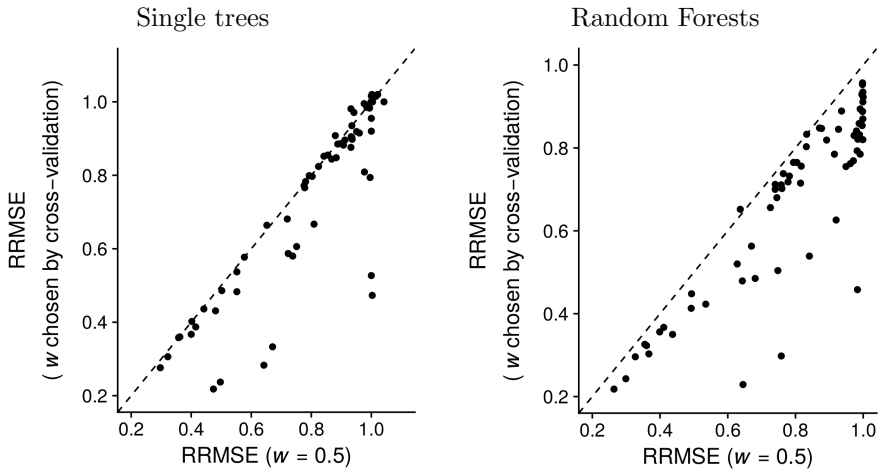


Figure 4: Comparison of the predictive performance of semi-supervised models induced with w chosen by internal 3-fold cross-validation and with $w = 0.5$, which weights equally labeled and unlabeled examples. Points below the diagonal line indicate that the later ($w = 0.5$) performs worse (higher RRMSE).

regression models that are easily interpretable – they are as easy to interpret as supervised regression trees. Therefore, as in common regression trees, the interpretability of the tree-shaped models is influenced by their size: a very large tree may be hard to analyze, and vice versa, a tree with fewer nodes may be easier to understand. In practice, the size of the tree is usually a trade-off between accuracy and interpretability. Namely, very small trees are easy to understand, but they may fail to capture structural information in the data and therefore fail to deliver satisfactory predictive performance. On the other hand, larger trees may avoid these issues, but at the cost of reduced understandability. Certainly, increased size of the tree does not necessarily imply improved predictive capability, due to the danger of overfitting. In general, it is hard to tell (*a priori*) when the tree is "large enough", i.e., whether it overfits or underifts.

Here, we discuss the differences in model sizes between the supervised and semi-supervised trees. The differences are given in Table 3. We can observe that semi-supervised regression trees are, on average, of a very similar size as

the supervised regression trees. This suggests that SSL-PCTs have a more favourable trade-off between accuracy and interpretability than CLUS-PCTs. In fact, SSL-PCTs offer comparable interpretability of CLUS-PCTs, but a better predictive performance.

Table 3: Size of regression trees expressed with number of nodes. The trees are learned by using the supervised (SL) and the semi-supervised (SSL) regression trees method.

Dataset	Number of labeled examples											
	25		50		100		200		350		500	
	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL	SL	SSL
2dplanes	4.4	10	9.6	18.8	17.8	32.2	32.8	47.2	50	50	61.4	61.4
Abalone	4	2	5.2	2.8	7.8	5.4	14	8	20.8	11.4	27.6	15.2
Elevators	4	1.6	4.6	2.8	13.6	4.8	19	19	29.4	10.2	36.8	15
Kinematics	3.2	1	5.6	3.2	5.8	5.4	10.4	7.8	18.8	22.2	29.8	17
Laser	5.4	6.4	11	11.4	21.2	16.6	39.4	23	87.4	87.4	106.2	106.2
Plastic	2.4	1	6.2	6.2	13	13	22.2	22.2	35.2	35.2	43	43
Pole	5.6	10	8.2	19.8	14.6	31.4	28.2	61.2	45.4	75.4	60.4	102
Stock	7.4	7.4	12.4	12.4	26	26	47.4	47.4	74.8	74.8	104	104
NK1	1.8	1.4	4.4	2	7	1	12.2	12.2	26.8	26.8	31.4	31.4
GSK3A	2.8	1	4	1	12.4	1	19.8	2.2	35.6	35.6	45.4	45.4
ROCK2	2.6	1	3.8	3.8	6.6	6.6	16.4	16.4	22.4	22.4	37.4	37.4
HIV-1	2.4	1.2	3.2	1	5	4.4	11.6	11.6	21.6	21.6	28.8	28.8
Average:	**3.8**	**3.7**	**6.5**	**7.1**	**12.6**	**12.3**	**22.8**	**23.2**	**39.0**	**39.4**	**51.0**	**50.6**

5. Use case: Predicting inhibitors of farnesyltransferase

The results reported in Figure 2 suggest that the proposed semi-supervised algorithms may be especially suitable for QSAR modeling, since on all four used datasets from this domain, either SSL-PCT or SSL-RF (or both methods) outperform their supervised counterparts. This observation is appealing, since the domain of QSAR modeling is particularly suitable for semi-supervised learning: Determining the biological activity of chemical compounds is a very expensive and tedious process (DiMasi et al., 2003), while descriptions of the structure of hundreds of thousands of unlabeled compounds are freely available in public

databases, such as ChEMBL (Bento et al., 2014).

Motivated by the above-mentioned observations, we pursue a practical application of the proposed semi-supervised method on the domain of QSAR modelling. The application is practical in a sense that we extract *real* unlabeled data from a public database and feed it to our algorithm. Note that the vast majority of work on semi-supervised learning published in the scientific literature actually simulates semi-supervised learning: Unlabeled data are simulated by sampling from labeled datasets and temporary removing their labels. This is somewhat contradictory to the motivation behind semi-supervised learning, i.e., the exploitation of unlabeled data that are easily and/or freely available in large quantities.

By applying the proposed semi-supervised methods we develop models for predicting inhibitors of *farnesyltransferase* (FTase). FTase is one of the three enzymes in the prenyltransferase group that catalyzes most prenylation reactions. Its targets include members of the Ras superfamily, which plays pivotal roles in control of cell growth. Ras genes are mutated in 30% of human cancers. Therefore, FTase inhibitors have been developed as anticancer drugs (Agrawal & Somani, 2009).

From the ChEMBL database (Bento et al., 2014), we extracted a dataset of 57 compounds for which inhibition of FTase was measured in the model organism *Saccharomyces cerevisiae S288c* (ID in the database: CHEMBL2111393), expressed as $-\log IC_{50}$ - the negative logarithm of the concentration of a compound causing 50% enzyme inhibition. Note that the sizes of datasets in the QSAR modelling domain commonly range from tens to hundreds of molecules.

We next extracted unlabeled compounds from ChEMBL, i.e., compounds for which their inhibitory property of FTase is unknown. To ensure that the unlabeled compounds bear structural similarity to labeled dataest, we queried for compounds with at least 0.8 Tanimoto similarity to the labeled compounds (Willett, 2006). This yielded 74 unlabeled compounds. We then trained supervised and semi-supervised trees and random forests, where semi-supervised algorithms were provided with unlabeled data in addition to labeled data. The

structures of the compounds were described with MACCS structural keys fingerprints (MACCS-II, 1984), calculated with the RDKit library (RDKit, 2018). The fingerprints are binary vectors of length 166, where each bit corresponds to a specific SMARTS[3] pattern.

Estimation of the predictive performance of the algorithms via 10-fold cross validation yielded RMSE of 0.764 and 0.683 for CLUS-PCT and SSL-PCT, respectively, and RMSE of 0.766 and 0.755 for CLUS-RF and SSL-RF, respectively. Therefore, by exploiting unlabeled data freely available in the ChEMBL database, the predictive performance of regression trees and random forests was improved by 10% and 1%, respectively.

Figure 5 depicts the regression trees obtained by the CLUS-PCT and SSL-PCT algorithms. It illustrates the interpretabiliy of the models obtained by this semi-supervised algorithm. We can see that both trees have the same size. They only differ in one split, and according to the evaluation of predictive performance, the SSL-PCT algorithm selects a better split.

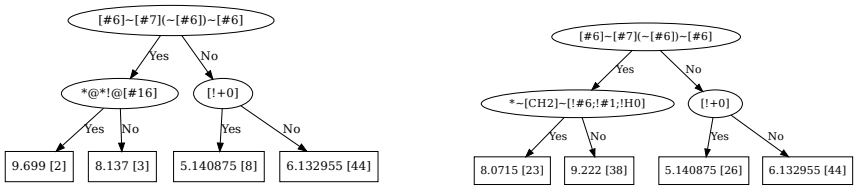
The patterns used in the trees in Figure 5 are graphically explained in Figure 6. The most important pattern denotes the presence of an NC_3 group. The pattern selected by SSL-PCT denotes the presence of a CH_2 group, connected to a non-C and non-H atoms further connected to some H atom (non-H_0).

6. Related work

In this section, we review several prominent semi-supervised regression methods and applications of semi-supervised learning to QSAR modeling. We refer to (Kostopoulos et al., 2018) for a more detailed and complete overview of semi-supervised regression methods.

One of the main approaches to semi-supervised regression relies on the multi-view learning approach, where two or more regressors are trained on different views of the data (Sindhwani et al., 2005; Brefeld et al., 2006; Zhou & Li, 2007;

[3]SMARTS is a line notation developed by Daylight Chemical Information Systems for compactly representing molecular substructure queries.



(a) (b)

Figure 5: The regression trees for predicting the inhibition of FTase, induced by the (a) supervised and (b) semi-supervised regression tree learning algorithms. The test nodes contains the SMARTS patterns that were searched for in the molecules. Figure 6 shows the explanations of the SMARTS patterns.

Kakade & Foster, 2007; Appice et al., 2010). Such methods rely on the multi-view assumption (Ceci et al., 2015), which is an extension of the co-training assumption to multiple views: Each view should be sufficient to build a model and features belonging to different views should be as independent as possible. This assumption, however, limits the applicability of such approaches since, in practice, different views complying with the multi-view assumption are often not available.

Belkin et al. (2006) proposed the Laplacian Regularized Least Squares (LapRLS) algorithm, which was long considered as a state-of-the-art semi-supervised regression algorithm. More recently, Ji et al. (2012) proposed the SSSL algorithm which finds top eigenfunctions spanning the feature space of labeled and unlabeled examples and then trains a model using labeled examples considering the subspace spanned by these eigenfunctions. The method was shown to outperform LapRLS, but is computationally very demanding, making scalability an issue. McWilliams et al. (2013) successfully dealt with the multi-view assumption by automatically constructing views from the data in such a way that they satisfy the assumption by design. Their method, named XNV, was shown to outperform SSSL, while being computationally efficient.

However, to the best of our knowledge, the existing semi-supervised regres-

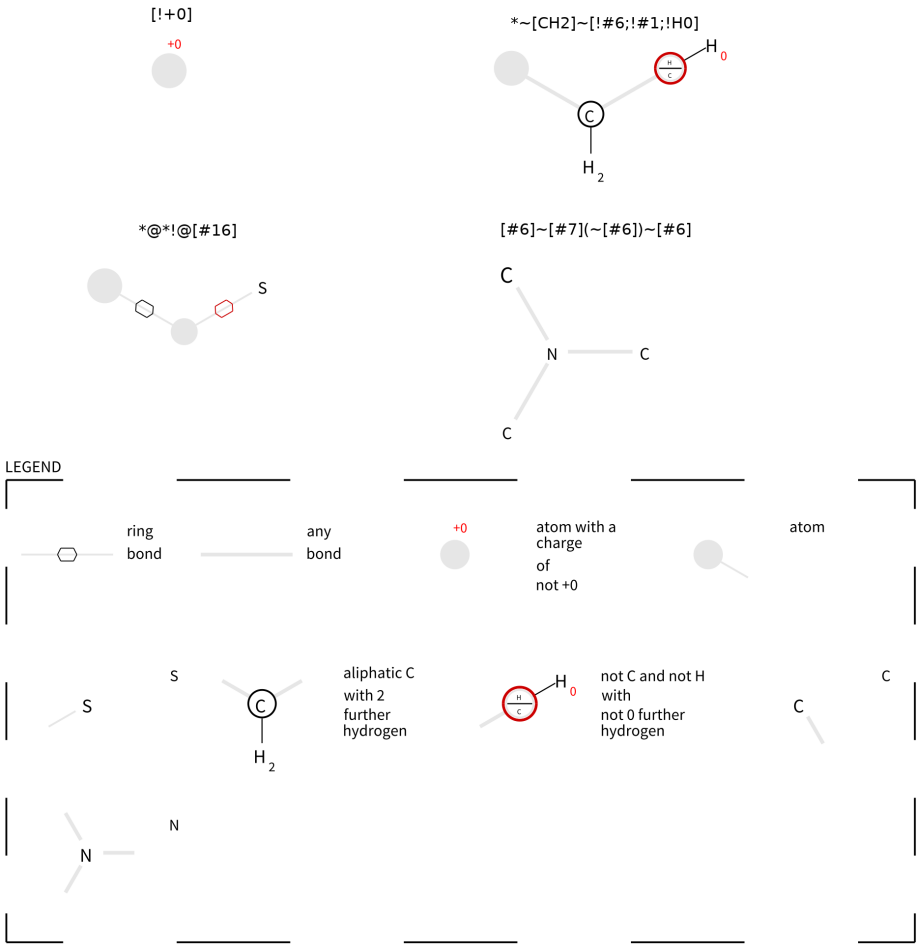


Figure 6: Visualization of the SMARTS patterns used in the trees in Figure 5. The image was generated with the help of *SMARTSviewer*, available at `https://smartsview.zbh.uni-hamburg.de`, ZBH Center for Bioinformatics, University of Hamburg.

sion algorithms (including the ones mentioned before) do not produce readily interpretable models. In principle, interpretability could be achieved by applying the self-training approach (Yarowsky, 1995) to regression trees. Self-training iteratively re-trains the model by using its own most reliable predictions on unlabeled data as additional training examples; therefore, if the underlying supervised method produces interpretable models, also the final model will be interpretable. Self-training has an inherent danger of error propagation: If a wrongly predicted example enters the training set, the error may propagate in subsequent iterations deteriorating the predictive performance of the model (Guo et al., 2010). For this reason, powerful regression methods are usually used as underlying methods of self-training, such as support vector regression (Kang et al., 2016) or random forests (Levatić et al., 2017). This study presents semi-supervised regression trees which are readily interpretable and follow a recent trend in semi-supervised learning: development of *safe* semi-supervised algorithms (Gan et al., 2018b,a; Li et al., 2017). Such algorithms are aimed to reduce the risk of unlabeled data degrading the performance and ideally should guarantee performance better than or as good as the one of supervised algorithms. This is achieved without using the self-training framework. The algorithms proposed in this study achieve this by implementing flexibility in terms of how much influence unlabeled data has during the learning process.

Closest to the work presented in this paper is our own work on semi-supervised multi-target regression. This includes the use of multi-target regression tree ensembles in a self-training setting (Levatić et al., 2017), as well as an approach similar to the one presented here (Levatić et al., 2018). The first of these two lines of work does not produce interpretable models, so we don't discuss it in more detail.

Levatić et al. (2018) does produce interpretable models, but focuses on multi-target, rather than single-target regression. No datasets from the domain of QSAR are used in the evaluation of (Levatić et al., 2018). Even more importantly, the conclusions regarding the experimental comparison differ substantially. In the case of multi-target regression, semi-supervised ensembles do not

perform significantly better than supervised ensembles, for any of the different amounts of labeled data. This is in stark contrast with the present paper, where for single-target regression, semi-supervised ensembles perform significantly better than the supervised ones, for almost all amounts of labeled data. The probable explanation for this lies in the fact that the better performance of semi-supervised learning is largely due to preventing overfitting (Levatić, 2017). Multi-target regression tree ensembles overfit less than their single-target counterparts, leaving less room for improvement by semi-supervised learning. Since regression tree ensembles overfit more, the additional unlabeled data improves their performance more noticeably.

Despite the fact that semi-supervised learning solutions seem to be, in principle, appropriate for QSAR modeling, due to difficulty of labeling the data and abundance of unlabeled data, very few published works combine semi-supervised learning and QSAR modeling (Guo-Zheng et al., 2008; Levatić et al., 2013; Kondratovich et al., 2013; Pan & Wei, 2012; Seeland et al., 2012). This study aims to contribute to both fields: semi-supervised learning and QSAR modeling, by proposing a novel semi-supervised regression method and demonstrating its advantage over supervised learning in QSAR modeling and other domains.

Furthermore, interpretability of QSAR models is an important property for their usability and practical acceptance (Cherkasov et al., 2014). Cherkasov et al. (2014) discusses approaches to indirectly interpret black-box QSAR models and hence faces the problem of simultaneously achieving both accurate and interpretable predictive models. Such approaches include the analysis of feature importance scores of molecular descriptors or the virtual modification of molecules by adding or removing fragments of interest, followed by re-doing the predictions in order to asses the importance of these fragments with respect to the activity of a compound. Our study presents semi-supervised regression trees which are natively interpretable, as well as semi-supervised random forests, which can offer state-of-the-art predictive performance when accuracy is a priority.

7. Conclusion

In this paper, we propose and evaluate a method for semi-supervised learning of regression trees and ensembles thereof. We have evaluated the method and its variants on a number of benchmark regression datasets and several datasets in the chemoinformatics domain of quantitative structure-activity relationship (QSAR) modeling. We have also performed a case study in QSAR modelling i.e., the determination of the biological activity of chemical compounds.

The proposed approach is motivated by the need to exploit, during the model learning stage, the vast amount of unlabeled data in addition to small sets of labeled data, which is particularly relevant in the QSAR domain. Unlabeled data, although not directly connected to labels (e.g., biological activity of compounds), can still convey useful information for learning models with better predictive performance. An additional advantage of the proposed approach is that it learns regression trees, which can be easily interpreted and understood by domain experts. This property is considered very important for QSAR modeling since it facilitates the understanding of the predictions being made.

Obviously, extending regression tree induction to the semi-supervised learning setting requires novel heuristic functions which take into account both the target and the descriptive spaces. For this reason, we extend predictive clustering trees – these naturally support extensions in such direction. The experiments confirm the expectations and prove the effectiveness of the semi-supervised learning approach in exploiting unlabeled examples in the induction of more accurate models. This aspect is clear both when we learn single regression trees and when we learn random forests. Moreover, a case study demonstrates the interpretability of extracted QSAR models. Experiments also prove the effectiveness of the proposed learning algorithm in other domains, different from QSAR modeling. This confirms the general-purpose nature of the proposed approach.

For future work, we first intend to study and develop transfer learning approaches to "transfer" the quantitative structure-activity relationship acquired

on some specific compounds to other (similar) compounds. Next, we will investigate other heuristic functions that take into account other impurity functions (e.g., entropy for discrete variables). Finally, we will use the proposed semi-supervised regression trees in order to provide a ranking of the feature, based on their relevance for the semi-supervised regression task.

Acknowledgements

We acknowledge the financial support of the Slovenian Research Agency, via the grants P2-0103 and J2-9230, and a young researcher grant to JL and TS, as well as the European Commission, via the grants ICT-2013-612944 MAESTRA, H2020-2018-785907 HBP SGA2 and H2020-ICT-688797 TOREADOR.

References

Agrawal, A. G., & Somani, R. R. (2009). Farnesyltransferase inhibitor as anti-cancer agent. *Mini Reviews in Medicinal Chemistry*, *9*, 638–652.

Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2010). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, *17*, 255–287.

Appice, A., Ceci, M., & Malerba, D. (2010). Transductive learning for spatial regression with co-training. In *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC)* (pp. 1065–1070). ACM Press.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*, 2399–2434.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., & Overington, J. P. (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Research*, *42*, D1083–D1090.

Blockeel, H., De Raedt, L., & Ramon, J. (1998). Top-down induction of clustering trees. In *Proceedings of the 15th International Conference on Machine learning* (pp. 55–63). Morgan Kaufmann.

Blockeel, H., & Struyf, J. (2002). Efficient algorithms for decision tree cross-validation. *Journal of Machine Learning Research*, *3*, 621–650.

Brefeld, U., Gärtner, T., Scheffer, T., & Wrobel, S. (2006). Efficient co-regularised least squares regression. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 137–144).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks.

Ceci, M., Pio, G., Kuzmanovski, V., & Džeroski, S. (2015). Semi-supervised multi-view learning for gene network reconstruction. *PLOS ONE*, *10*, 1–27.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised Learning*. MIT Press.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuzmin, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, *57*, 4977–5010.

Cozman, F., Cohen, I., & Cirelo, M. (2002). Unlabeled data can degrade classification performance of generative classifiers. In *Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference* (pp. 327–331).

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, *22*, 151–185.

Escalante, H. J., Guyon, I., Escalera, S., Jacques, J. C. S., Madadi, M., Baró, X., Ayache, S., Viegas, E., Güçlütürk, Y., Güçlü, U., van Gerven, M. A. J., & van Lier, R. (2017). Design of an explainable machine learning challenge for video interviews. In *IJCNN* (pp. 3688–3695). IEEE.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). *Foundations of Rule Learning*. Springer.

Gan, H., Li, Z., Fan, Y., & Luo, Z. (2018a). Dual learning-based safe semi-supervised learning. *IEEE Access*, *6*, 2615–2621.

Gan, H., Li, Z., Wu, W., Luo, Z., & Huang, R. (2018b). Safety-aware graph-based semi-supervised learning. *Expert Systems with Applications*, *107*, 243–254.

Guo, Y., Niu, X., & Zhang, H. (2010). An extensive empirical study on semi-supervised learning. In *Proceedings of the 10th International Conference on Data Mining* (pp. 186–195).

Guo-Zheng, L., Jack, Y. Y., Wen-Cong, L., Dan, L., & Mary, Q. Y. (2008). Improving prediction accuracy of drug activities by utilizing unlabeled instances with feature selection. *International Journal of Computational Biology and Drug Design*, *1*, 1–13.

Ji, M., Yang, T., Lin, B., Jin, R., & Han, J. (2012). A simple algorithm for semi-supervised learning with improved generalization error bound. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1223–1230).

Kakade, S. M., & Foster, D. P. (2007). Multi-view regression via canonical correlation analysis. In *Proceedings of the 20th Conference on Learning Theory, LNCS, vol. 4539* (pp. 82–96). Springer.

Kang, P., Kim, D., & Cho, S. (2016). Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. *Expert Systems with Applications*, *51*, 85–106.

Kocev, D., Vens, C., Struyf, J., & Džeroski, S. (2013). Tree ensembles for predicting structured outputs. *Pattern Recognition*, *46*, 817–833.

Kondratovich, E., Baskin, I. I., & Varnek, A. (2013). Transductive support vector machines: Promising approach to model small and unbalanced datasets. *Molecular Informatics*, *32*, 261–266.

Kostopoulos, G., Karlos, S., Kotsiantis, S., & Ragos, O. (2018). Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, *35*, 1483–1500.

Levatić, J. (2017). *Semi-Supervised Learning for Structured Output Prediction*. Ph.D. thesis Jožef Stefan International Postgraduate School Ljubljana, Slovenia.

Levatić, J., Ceci, M., Kocev, D., & Džeroski, S. (2017). Self-training for multi-target regression with tree ensembles. *Knowledge-Based Systems*, *123*, 41–60.

Levatić, J., Džeroski, S., Supek, F., & Šmuc, T. (2013). Semi-supervised learning for quantitative structure-activity modeling. *Informatica (Slovenia)*, *37*, 173–179.

Levatić, J., Kocev, D., Ceci, M., & Džeroski, S. (2018). Semi-supervised trees for multi-target regression. *Information Sciences*, *450*, 109 – 127.

Li, Y.-F., Zha, H.-W., & Zhou, Z.-H. (2017). Learning safe prediction for semi-supervised regression. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 2217–2223).

MACCS-II (1984). MDL Information Systems/Symyx, Santa Clara, CA.

Malerba, D., Ceci, M., & Appice, A. (2009). A relational approach to probabilistic classification in a transductive setting. *Eng. Appl. of AI*, *22*, 109–116. URL: `https://doi.org/10.1016/j.engappai.2008.04.005`. doi:`10.1016/j.engappai.2008.04.005`.

Malerba, D., Esposito, F., Ceci, M., & Appice, A. (2004). Top-down induction of model trees with regression and splitting nodes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*, 612–625.

McWilliams, B., Balduzzi, D., & Buhmann, J. M. (2013). Correlated random features for fast semi-supervised learning. In *Proceedings of the 23rd Conference on Advances in Neural Information Processing Systems* (pp. 440–448). Curran Associates, Inc.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*, 103–134.

Olier, I., Sadawi, N., Bickerton, G. R., Vanschoren, J., Grosan, C., Soldatova, L., & King, R. D. (2018). Meta-qsar: a large-scale application of meta-learning to drug design and discovery. *Machine Learning*, *107*, 285–311.

Pan, Z., & Wei, X. (2012). A graph based transductive ranking algorithm. In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 991–994). IEEE.

Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning*. (1st ed.). Morgan Kaufmann.

RDKit (2018). RDKit, Open-Source Chemoinformatics Software. `https://github.com/rdkit/rdkit`. Version: 2018.3.1.

Seeland, M., Karwath, A., & Kramer, S. (2012). A structural cluster kernel for learning on graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 516–524). ACM.

Sindhwani, V., Niyogi, P., & Belkin, M. (2005). A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views* (pp. 74–79).

Torgo, L. (2016). Luís Torgo - Regression data sets. URL: `http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html`.

Vanschoren, J., Van Rijn, J. N., Bischl, B., & Torgo, L. (2014). OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, *15*, 49–60.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, *11*, 1046–1053.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pp. 189–196). Association for Computational Linguistics Stroudsburg, PA.

Zhou, Z.-H., & Li, M. (2007). Semi-supervised regression with co-training style algorithms. *IEEE Transaction on Knowledge and Data Engineering*, *19*, 1479–1493.

Zhu, X. (2008). *Semi-Supervised Learning Literature Survey*. Technical Report Computer Sciences, University of Wisconsin-Madison.