Semi-Supervised Learning for Quantitative Structure-Activity Modeling

Jurica Levatić
Faculty of Science, Department of Mathematics, University of Zagreb, Zagreb, Croatia
Bijenička cesta 30, 10000 Zagreb, Croatia
E-mail: jurica.levatic@ijs.si

Sašo Džeroski
Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: saso.dzeroski@ijs.si

Fran Supek and Tomislav Šmuc
Division of Electronics, Institute Ruđer Bošković, Zagreb, Croatia
Bijenička cesta 54, 10000 Zagreb, Croatia
E-mail: fran.supek@irb.hr, smuc@irb.hr

Keywords: semi-supervised learning, supervised learning, QSAR, drug design, machine learning

Received: September 6, 2012

In this study, we compare the performance of semi-supervised and supervised machine learning methods applied to various problems of modeling Quantitative Structure Activity Relationship (QSAR) in sets of chemical compounds. Semi-supervised learning utilizes unlabeled data in addition to labeled data with the goal of building better predictive models than can be learned by using labeled data alone. Typically, labeled QSAR datasets contain tens to hundreds of compounds, while unlabeled data are easily accessible via public databases containing thousands of chemical compounds: this makes QSAR modeling an attractive domain for the application of semi-supervised learning. We tested four different semi-supervised learning algorithms on three different datasets and compared them to five commonly used supervised learning algorithms. While adding unlabeled data does help for certain pairings of dataset and method, semi-supervised learning is not clearly superior to supervised learning across the QSAR classification problems addressed by this study.

Povzetek: Metode delno-nadzorovanega učenja smo testirali na različnih podatkih iz domene kvantitativnega modeliranja razmerja med strukturo in aktivnostjo kemičnih spojin (angl. Quantitative Structure Activity Relationship, oziroma QSAR).

1 Introduction

Two major approaches to machine learning are supervised learning (e.g., classification, regression), where all the data are labeled, and unsupervised learning (e.g., clustering, dimensionality reduction) where all the data are unlabeled. The semi-supervised learning (SSL) paradigm [21] examines how merging both types of data (labeled and unlabeled) affects learning, aiming to benefit from the information that unlabeled data bring in the context of the supervised learning tasks.

SSL is of important practical value since the following scenario often holds true: labeled data are scarce and hard to get because they require human experts, expensive devices or time-consuming experiments, while, at the same time, unlabeled data abound and are easily obtainable. Real-world classification problems of this type include: phonetic annotation of human speech, protein 3D structure prediction, and spam filtering. Intuitively, SSL yields best results when there are few labeled examples as compared to unlabeled ones (i.e., large-scale labelling is not affordable). But, the setting where plenty of labeled data are available is also suitable for SSL, if even more unlabeled data are available. The other scenario where SSL can be applied is ‘domain adaptation’; where we have labeled examples belonging to one domain, but we want to develop a model for another, related, domain.

Establishing a connection between biological effects and structural and/or physicochemical properties of chemicals is the task of quantitative structure-activity relationship or QSAR modeling. Formal studies of such relationships are the basis for the development of predictive models. The main value of a predictive QSAR model is the fact that it provides insight into the biological activity of a molecule without the need to synthesize it. This leads to a number of benefits including savings in the cost and duration of product development (e.g., in the pharmaceutical or pesticide industries), reduction of the need for animal testing, prediction of unwelcome or toxic environmental impact, and overall improvement in the efficiency of drug design.

The application of SSL to the domain of QSAR modeling is particularly attractive since the premise: “labeled data are scarce, while unlabeled data abound” is generally satisfied in this domain. Public databases with (hundreds of) thousands of chemical compounds are available (e.g., the human tumor cell line screen database from the U.S. National Cancer Institute’s Developmental Therapeutics program), while labeled datasets sizes typically range from tens to hundreds and rarely surpass a thousand molecules.

In this work, we empirically investigate whether we can successfully apply SSL (i.e., whether we can achieve better performance with SSL than with supervised learning) to build predictive QSAR models. To draw reliable conclusions, we use several SSL methods which embody different approaches, together with three QSAR datasets from various domains. We compare the SSL methods to several commonly used supervised learning methods. The results show that the improvements which SSL yields are selective - the degree to which unlabeled data help varies from notable to insignificant, depending on the dataset or SSL method used.

2 Semi-supervised learning

In this study, we are concerned with semi-supervised classification, while other forms of SSL, such as semi-supervised regression or semi-supervised clustering are not considered.

2.1 The task of semi-supervised classification

In supervised learning, we are given training data in the form of instance-label pairs, i.e., for each instance we know the desired prediction. The goal is to use the training data to infer a mapping, from instances to labels, which will provide (true) labels for future instances. If the domain of labels is discrete, such a mapping is called a classification function (or a classifier).

The task of *semi-supervised classification* is an extension to the task of supervised classification, where the training data, in addition to the labeled instances, contain a set of unlabeled instances. The goal is again to produce a classification function, which hopefully performs better than the classifier learned from the supervised data only classifier. Figure 1 shows a simple example how unlabeled data can help to induce a classifier that is better in separating the classes.

2.2 Major approaches to semi-supervised classification

In order for SSL to work, the knowledge we gain trough unlabeled data has to carry some information about the class labels. If this prerequisite is fulfilled, we can draw on unlabeled data by making certain assumptions about the behavior of labels with respect to the structure of unlabeled data. Different assumptions inspire different classes of algorithms; therefore, SSL methods can be grouped on the basis of the assumption(s) they implement as follows: low-density separation methods, graph-based methods, generative models, self-training and co-training.

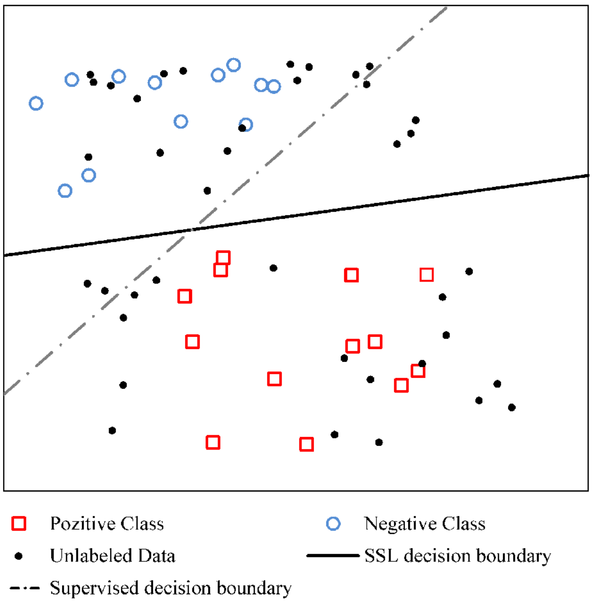

Figure 1. Semi-supervised linear SVMs use unlabeled data to find decision boundary which separates the two classes better than the decision boundary discovered by supervised SVMs.

Low-density separation methods assume that the decision boundary should lie in the region of low density of the data. For example, semi-supervised support vector machines try to find a labeling for the unlabeled data in a way that maximizes the margin of the decision boundary considering both labeled and unlabeled data. Equivalent to the low density separation assumption is *the cluster assumption*: the points belonging to the same cluster should be of the same class.

Graph-based methods use nodes for data representation (labeled and unlabeled) and edges (usually with weights representing the similarity of the data points) for propagation of the labels through the graph, assuming label smoothness over the graph (i.e, the label of the unlabeled instance should be similar to its neighbors in the graph). Here, unlabeled data help to “bridge” the points which would otherwise be unconnected. The construction of the graph is a critical step of graph-based methods – it should reflect the information which is not easily encoded in feature vectors.

Generative models assume a probabilistic model of the data and use unlabeled, together with labeled data, to estimate the most probable model parameters. The success of generative models depends largely on choosing a probabilistic model which is appropriate for the data. Once the probabilistic model is chosen (e.g., Gaussian mixture models), a maximum likelihood estimate (MLE) of the parameters can be calculated (e.g., by using the Expectation-Maximization algorithm), followed by a calculation of class distributions using Bayes' rule.

Self-training and *co-training* are two approaches that are often used by SSL algorithms, since they can be "wrapped" around any (supervised) learning algorithm. They iteratively use their own most reliable predictions in the training process (assuming they are correct), as additional data for learning. The main pitfall of these methods is the reinforcement of mistakes – a mistake once made can reinforce itself in the next iterations, leading to degradation of performance.

These assumptions are at the heart of SSL, but also present the main risk for bad performance of SSL: an inappropriate match of a problem structure to a method's assumption can cause severe degradation of performance when using unlabeled data [2]. This is a particularly relevant issue since it is not yet clearly understood which SSL method should be used for which problem, or whether a certain problem (or dataset) is suitable for SSL the use of at all. As mentioned before, unlabeled data has to carry useful information about the structure of the data with respect to the labels.

Zhang and Oles [19] tried to quantify the value of unlabeled data in a probabilistic framework by using regularized logistic regression as an approximation of support vector machines. They showed that, in the setting where labeled and unlabeled data do not share parameters, semi-supervised support vector machines are unlikely to be helpful in general, and are prone to maximize the "wrong margin". It should be noted that unlabeled data should not be used to compensate for the lack of labeled data, but to complement labeled data. In other words, the improvements based on SSL should not rely on the inability of supervised methods to learn anything useful at all due to the lack of data.

We tackled the difficulties of matching the problem structure with the right SSL method empirically, i.e., by selecting methods which differ in their basic approach. We tried to cover most of the groups of methods mentioned above. The SSL methods we used will be described in Section 4.

3 QSAR datasets

To better assess the performance of SSL algorithms in the domain of QSAR modeling, we extracted three different datasets from publicly available sources. These are the NCI, Mutagenicity and MUSK dataset. The datasets differ in terms of the biological activity they model, the number and type of molecular descriptors used to represent molecules, and the number of compounds (size of the dataset).

3.1 NCI dataset

The NCI datasets was extracted from the human tumor cell line screen database [11] of the National Cancer Institute's Developmental Therapeutics (NCI-DTP) program (October 2009 release). The NCI-DTP measures cytostatic activity of chemical compounds against 60 human tumor cell lines grown in cell culture. For representation of a compound's cytostatic activity we used GI_{50} measurements – the compound concentration that inhibits cell growth by 50%. Only compounds that have missing or default values for at most 20 cell lines were accepted. Additionally, cell lines with more than 20% of missing values were removed, leaving 49 cell lines in total. The compounds were thus described with the GI_{50} profiles across the 49 cell lines, and in addition with two other groups of attributes: (1) molecular descriptors describing the structure of a molecule (calculated with the DRAGON 3.0 web interface [18]), and (2) molecular charge densities and charge density-based electrostatic properties of a molecule (calculated with the RECON software [4]).

The subject of interest for the NCI dataset is to predict a compound's mechanism of action (MOA) – the biological process in which the molecule interacts with its molecular targets - proteins (enzymes or otherwise) or DNA. The type of MOA influences the pharmacological effects of a molecule; therefore, the drug discovery process benefits from an early detection of an appropriate MOA for a given use. The NCI dataset represents a multiclass classification problem, with 12 different MOA classes, where each molecule belongs to a single class. A very similar dataset has been used to find putative MOAs for new drug candidates [7, 16], and is essentially an updated and extended version of the dataset used in previous analyses of cytostatic activities and MOA in global computational analyses of the NCI database using self-organizing maps [13, 15].

3.2 Mutagenicity dataset

The Mutagenicity dataset [10] is the benchmark dataset for modeling of Ames mutagenicity. The Ames test is a standard microbiological assay for assessing the mutagenic potential of a chemical compound. A compound which is positive to the test causes mutations on the DNA (and consequently can be carcinogenic); avoiding mutagenicity is important for drug-candidates and other molecules with significant human exposure (e.g., cosmetics, food additives).

The mutagenicity dataset represents a binary classification problem where compounds are classified as mutagenic or non-mutagenic. Molecules from this dataset were represented by using DRAGON molecular descriptors [18].

3.3 MUSK dataset

The MUSK dataset was downloaded from the UCI machine learning repository [8]. Musk, a substance secreted by the Asian musk deer, is an expensive animal product heavily used by the perfume industry; therefore, synthetic compounds are often used instead. The prediction of the strength of such synthetic musk compounds has similarities to the prediction of biological drug activity – the molecules are similar in size and composition to the orally active drug molecules [5].

A single molecule can adopt multiple conformations – different shapes of the same molecule, when some of the internal bonds rotate. The features that describe compounds from the MUSK dataset depend on the exact shape (conformation) of a molecule ("distance features" and displacement of oxygen; a detailed description is given by Dietterich et al. [5]), where each molecule is represented by several feature vectors. This dataset was assembled by generating low-energy conformations of molecules, which were then filtered to remove highly similar conformations. The molecules from the MUSK dataset were categorized by human experts to be musk or non-musk.

4 Experimental setup

To evaluate the potential of SSL in a controlled manner (i.e., to be able to evaluate the methods thoroughly, and to make sure that the unlabeled data is relevant to the problem), our experiments were carried out using only labeled data. We simulated unlabeled data by temporarily ignoring the class label for a portion of the data. The relative amount of unlabeled and labeled data is a relevant factor when measuring the success of SSL methods: SSL should perform better when the labeled set is rather small and a lot of unlabeled data are available. Our experiments were aimed to test the former premise by creating situations where we have different ratios of labeled and unlabeled data.

The data were randomly split into a training and a test set. Both the supervised and the semi-supervised methods used the training set for learning and were then evaluated by using the test set. For the SSL methods, the test set served as unlabeled data during the learning process. Several different train/test splits were produced where labeled data ranges from 1% to 66% (i.e., unlabeled data ranges from 99% to 33%). The final results were averaged over 10 different train/test split repetitions, in order to obtain a more robust evaluation of the algorithms. We performed experiments using the Weka [9] machine learning environment and the R [17] environment for statistical computing.

4.1 Datasets

As described in Section 3, we conducted experiments on three different QSAR datasets. The NCI dataset contains 507 compounds, each described with: GI50 profiles (49 features in the form of -logGI50), DRAGON descriptors (1497 features) and RECON descriptors (248 features). The Mutagenicity dataset is the largest with 6512 compounds represented with 1497 DRAGON descriptors. The MUSK dataset has 166 features and 476 examples, which correspond to different conformations of 92 molecules.

4.2 Methods

We used publicly available implementations of several SSL algorithms. As mentioned in Section 2, we selected the SSL algorithms to cover different groups of SSL methods. The algorithms used are: Yet Another Two Stage Idea (YATSI), Co-training: Fitting the Fits (Co-FTF), Learning with Local and Global Consistency (LLGC) and TSVMLight.

The YATSI [6] algorithm, implemented in the Weka Collective Classifiers package, is similar to the self-training concept, since it can be wrapped around any classifier and it uses its own predictions in the training process. As the name implies, YATSI works in two steps. First, a base classifier is trained on the labeled data and then unlabeled data is "pre-labeled". This pre-labeled data is then given weights and used by the nearest neighbors classifier to improve on the initial classifier.

Co-FTF [3] is an implementation of the co-training algorithm in the R programing language. Co-FTF uses two different features sets (views) to train separate classifiers, which iteratively use their most confident predictions as additional labeled training data. It is assumed that views provide different, complementary information about the data. We applied Co-FTF only to the NCI dataset (the other datasets do not meet the prerequisite for different views) with the combination of the descriptors which proved to be the best: RECON and DRAGON. Other combinations: GI50 profiles coupled with RECON or DRAGON descriptors, achieved lower performances (not shown). The baseline classifier for Co-FTF was the random forests classifier with 500 trees.

LLGC [20] is a graph-based method implemented in the Weka Collective Classifiers package. LLGC first performs spectral clustering and then propagates labels through the graph using a spreading activation network.

TSVMLight [12] is a representative of the low-density separation methods. It implements a semi-supervised version of support vector machines by finding the locally optimal solution.

The supervised machine learning methods that we compared with SSL methods were taken from Weka: decision trees (J48), k-nearest neighbors (KNN), Naive Bayes (NB), support vector machines (SMO from Weka, and the stand-alone version of SVMLight) and random forests (RF).

We used the J48, NB and SMO methods with their default parameters and RF with 500 trees. For the KNN method, the 'crossValidate' option was used to select an appropriate number K of neighbours. For YATSI and LLGC, we used the Weka Experimenter Environment to search for the parameter values which produce the best classification accuracy. The parameters for (T)SVMLight were tuned manually.

5 Results and discussion

In this section we present the experimental comparison of performance of semi-supervised and supervised machine learning methods. In Tables 1-3, the predictive accuracies for different ratios of labeled and unlabeled data are presented. The best result for each ratio is shown in bold, and whether YATSI exhibited improvement in accuracy over the baseline classifier is marked with an upward (improvement) or downward (deterioration) arrow. The baseline classifier for YATSI is given in brackets. The number of neighbors for the KNN algorithm is indicated (e.g., 1NN).

Semi-supervised methods behave differently over the three datasets. Improvements of semi-supervised over supervised learning are most notable for the NCI dataset with a small percentage of labeled data (≤10%), where LLGC achieves the best overall predictive accuracy and YATSI significantly improves the baseline classifier in most cases. YATSI consistently deteriorates the performance of SMO for all amounts of labeled data.

For the other two datasets, Mutagenicity and MUSK, semi-supervised and supervised algorithms show very similar performance with small improvements of SSL over supervised learning in some cases. Generally, the improvements achieved by YATSI over the baseline classifier are more frequent and significant for the less complex classifiers (KNN, J48, NB), while for classifiers with greater capacity for learning (RF, SMO) the improvements are not so regular and are sometimes negative, i.e., the usage of YATSI even deteriorates their predictive accuracy (Figure 1).

Driessens et al. [6] performed an extensive testing of YATSI over 29 different datasets with several different base classifiers and made similar observations: YATSI behaves somewhat differently when using RF and SMO as base classifiers, as compared to the other algorithms (including J48 and KNN). In most cases, YATSI lost some of the accuracy achieved by RF, and performed equal to SMO, while it improved other base classifiers (with most notable improvements when little labeled data were available).

In the setting of supervised learning, robust methods, such as support vector machines or random forests are known to perform well on a wide range of classification tasks, and can be successfully used without specific domain knowledge. The results obtained on the datasets considered in this study confirm this: the SMO and RF

	Algorithm	Percentage of labeled data				
		5%	10%	20%	33%	66%
Supervised learning	J48	45.93	57.98	69.05	76.31	81.92
	1NN	47.45	67.19	78.14	82.73	86.80
	NB	42.41	51.19	66.13	74.49	84.14
	SMO	62.80	73.15	**83.42**	**87.41**	**92.69**
	RF	56.24	66.32	78.29	84.14	88.64
Semi-supervised learning	YATSI(J48)	55.37↗	68.27↗	78.54↗	83.31↗	84.87↗
	YATSI(1NN)	58.87↗	70.89↗	79.95↗	82.79↗	86.50↘
	YATSI(NB)	54.70↗	65.96↗	75.61↗	81.67↗	83.03↘
	YATSI(SMO)	62.06↘	72.69↘	81.99↘	84.76↘	87.59↘
	YATSI(RF)	58.76↗	68.44↗	79.11↗	83.46↘	86.32↘
	LLGC	**66.50**	**74.95**	82.46	85.46	88.29
	Co-FTF	-	35.78	51.16	65.43	76.45

Table 1: Predictive accuracies of semi-supervised and supervised learning methods on the NCI dataset

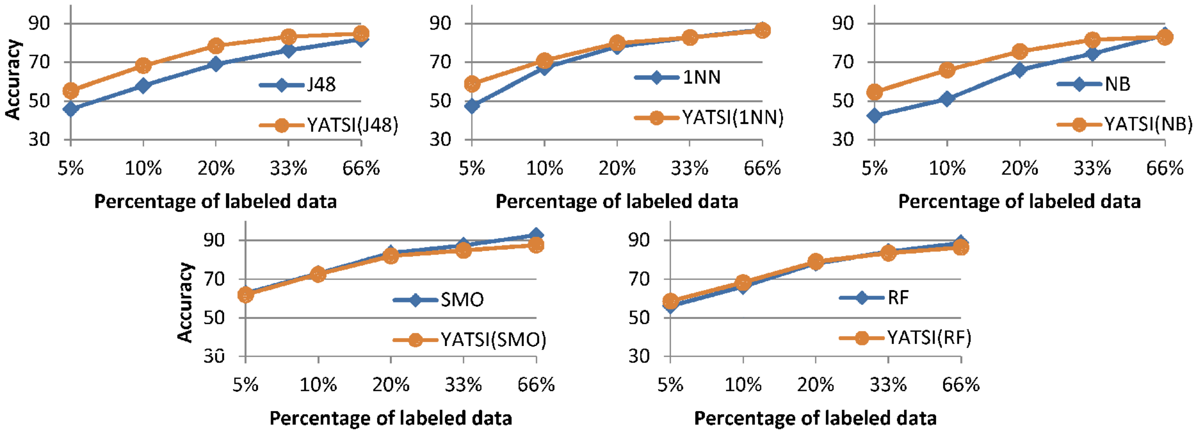

Figure 1. Comparison of learning curves for YATSI and baseline algorithms on the NCI dataset show that YATSI improves the performance of the less complex classifiers (J48, KNN, NB), but not the more complex classifiers (SMO and RF). The improvements in accuracy which unlabeled data brings gradually decrease with the increase of the relative amount of labeled data.

	Algorithm	Percentage of labeled data			
		1%	5%	10%	20%
Supervised learning	J48	58.40	63.98	67.08	69.97
	1NN	58.32	64.24	64.71	67.86
	NB	57.80	60.40	61.10	60.79
	SMO	61.86	68.95	72.41	**75.41**
	RF	62.13	68.68	71.27	73.68
	SVM$^{\text{Light}}$	**62.73**	69.29	72.52	75.16
Semi-supervised learning	YATSI(J48)	58.85↗	65.78↗	68.19↗	70.88↗
	YATSI(1NN)	58.45↗	64.57↗	66.77↗	69.30↗
	YATSI(NB)	57.85↗	62.71↗	64.62↗	65.02↗
	YATSI(SMO)	61.53↘	67.84↘	70.35↘	72.73↘
	YATSI(RF)	59.50↘	66.36↘	67.89↘	70.62↘
	TSVM$^{\text{Light}}$	61.24	**69.65**	**72.85**	**75.41**
	LLGC	58.75	62.70	63.65	64.86

Table 2: Predictive accuracies of semi-supervised and supervised learning methods on the Mutagenicity dataset

	Algorithm	Percentage of labeled data		
		5%	10%	20%
Supervised learning	2NN	63.89	71.15	77.97
	SMO	66.01	71.65	76.03
	RF	62.70	71.84	78.77
	SVM$^{\text{Light}}$	**69.49**	75.18	**81.02**
Semi-supervised learning	YATSI(2NN)	65.12↗	71.51↗	78.05↗
	YATSI(SMO)	67.83↗	74.42↗	78.07↗
	YATSI(RF)	63.57↗	73.25↗	78.48↘
	TSVM$^{\text{Light}}$	66.69	**75.25**	80.50
	LLGC	65.34	73.11	80.39

Table 3: Predictive accuracies of semi-supervised and supervised learning methods on the MUSK dataset.

classifiers consistently outperform the other (supervised) methods. However, if we compare SSL methods across the three datasets (Tables 1-3) we do not have a clear winner. For example, the LLGC algorithm performs better than the other SSL methods on the NCI dataset, but it is outperformed on the Mutagenicity and MUSK datasets.

Similar observations have been made by other scientists: Chawla and Karakoulas [1] performed an extensive empirical study of SSL techniques over various domains (not including QSAR modeling), using real-world and artificial datasets to investigate the conditions under which SSL can perform well. They observed that SSL methods behave very differently depending on the nature of the datasets, and that no single SSL method consistently performs better than supervised learning.

In practice, it is not easy to assess in advance how certain SSL method will behave given the task at hand. Several method/problem combinations are known to work well together (e.g., semi-supervised SVMs and text classification, [12]), but there are no clear strategies how to verify the model assumptions against certain problem structure. Specific domain knowledge and understanding of SSL algorithms should be used to couple the problem at hand with an appropriate method. Currently, scientists in this area are dealing with the question of how to make SSL safe, i.e., how to make sure that SSL performs at least as well as supervised learning, and how to make SSL usable by non-experts on realistic tasks [14].

6 Conclusion and future work

In this study, we performed an empirical comparison of several semi-supervised and supervised machine learning methods on three different QSAR datasets under different experimental conditions (amount of unlabeled data relative to labeled data). Our results show that SSL can achieve better predictive performance than supervised learning (typically when a small portion of the data is labeled), but the improvements depend on the dataset and method used. We cannot claim clear superiority of semi-supervised over supervised learning on the QSAR classification problems addressed by this study. However, the large improvements (in general and relative to the baseline classifier) in classification accuracy in certain cases suggest that it is worthwhile to take SSL into consideration when dealing with problems of QSAR modeling.

Semi-supervised learning is a more delicate task than supervised learning, where more (labeled) data generally a means better and more robust model. While more unlabeled data can help, it is not guaranteed to do so. We have pointed out the difficulties that one can encounter when dealing with the task of semi-supervised learning, as compared to supervised learning.

In further work, we would like to systematically investigate which features of a dataset make it suitable for the use of SSL. In addition, we would like to extend our experiments and use data which are truly unlabeled. This would enable us to exploit the vast amount of information readily available within public compound databases.

References

[1] Chawla, N.V. and Karakoulas, G. 2005. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*. 23 (1), 331–366.

[2] Cozman, F.G. et al. 2002. Unlabeled data can degrade classification performance of generative classifiers. In *Proc of the Fifteenth International Florida Artificial Intelligence Research Society Conference* (2002), 327–331. AAAI Press.

[3] Culp, M. and Michailidis, G. 2009. A co-training algorithm for multi-view data with applications in data fusion. *Journal of Chemometrics*. 23 (6), 294–303.

[4] Curt M. Breneman et al. 2003. *RECON version 5.5/5.3*. Rensselaer Polytechnic Institute.

[5] Dietterich, T.G. et al. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*. 89 (1–2), 31–71.

[6] Driessens, K. et al. 2006. Using weighted nearest neighbor to benefit from unlabeled data. In *Proc of the Knowledge Discovery and Data Mining*, 60–69. Springer.

[7] Ester, K. et al. 2012. Putative mechanisms of antitumor activity of cyano-substituted heteroaryles in HeLa cells. *Investigational New Drugs*. 30 (2), 450–467.

[8] Frank, A. and Asuncion, A. 2010. *UCI Machine Learning Repository:* University of California, Irvine, School of Information and Computer Sciences. *http://archive.ics.uci.edu/ml.* Accessed: January, 2011.

[9] Hall, M. et al. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. 11 (1), 10–18.

[10] Hansen, K. et al. 2009. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *Journal of Chemical Information and Modeling*. 49 (9), 2077–2081.

[11] Holbeck, S.L. 2004. Update on NCI in vitro drug screen utilities. *European Journal of Cancer*. 40 (6), 785–793.

[12] Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proc of the Sixteenth International Conference on Machine Learning*, 200–209. Morgan Kaufmann.

[13] Rabow, A.A. et al. 2002. Mining the National Cancer Institute's Tumor-Screening Database: Identification of Compounds with Similar Cellular Activities. *Journal of Medicinal Chemistry*. 45 (4), 818–840.

[14] Xiaojin Zhu, Semi-Supervised Learning for Non-Experts: *http://pages.cs.wisc.edu/~jerryzhu/ssl/.* Accessed: August, 2012.

[15] Supek, F. et al. 2005. A prototype structure-activity relationship model based on National Cancer Institute cell line screening data. *Periodicum Biologorum*. 107 (4), 451.

[16] Supek, F. et al. 2008. Atypical cytostatic mechanism of N-1-sulfonylcytosine derivatives determined by in vitro screening and computational analysis. *Investigational New Drugs*. 26 (2), 97–110.

[17] Team, R.C. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing

[18] Tetko, I.V. et al. 2005. Virtual computational chemistry laboratory--design and description. *Journal of Computer-aided Molecular Design*. 19 (6), 453–463.

[19] Zhang, T. and Oles, F. 2000. A probability analysis on the value of unlabeled data for classification problems. In *Proc of the Seventeenth International Conference on Machine Learning*, 1191–1198. Morgan Kaufmann.

[20] Zhou, D. et al. 2004. Learning with local and global consistency. *Advances in Neural Information Processing Systems*. 16, 321–328.

[21] Zhu, X. and Goldberg, A.B. 2009. *Introduction to Semi-Supervised Learning*. Morgan and Claypool.