

Conformal Prediction in Learning Under Privileged Information Paradigm with Applications in Drug Discovery

Niharika Gauraha

NIHARIKA.GAURAH@FARMBIO.UU.SE

*Department of Pharmaceutical Biosciences
Uppsala University
Uppsala, Sweden*

Lars Carlsson

LARS.A.CARLSSON@ASTRAZENECA.COM

*Quantitative Biology, Discovery Sciences,
IMED Biotech Unit, AstraZeneca
Gothenberg, Sweden*

*Department of Computer Science,
Royal Holloway, University of London,
Egham Hill, Egham, Surrey, United Kingdom*

Ola Spjuth

OLA.SPJUTH@FARMBIO.UU.SE

*Department of Pharmaceutical Biosciences
Uppsala University
Uppsala, Sweden*

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgeni Smirnov and Ralf Peeters

Abstract

This paper explores conformal prediction in the learning under privileged information (LUPI) paradigm. We use the SVM+ realization of LUPI in an inductive conformal predictor, and apply it to the MNIST benchmark dataset and three datasets in drug discovery. The results show that using privileged information produces valid models and improves efficiency compared to standard SVM, however the improvement varies between the tested datasets and is not substantial in the drug discovery applications. More importantly, using SVM+ in a conformal prediction framework enables valid prediction intervals at specified significance levels.

Keywords: Learning Under Privileged Information, LUPI, SVM, SVM+, conformal prediction, drug discovery

1. Introduction

The growing availability of data offers great opportunities but also many challenges to develop models which can be used to make predictions about future observations. The classical machine learning paradigm is: given a set of training examples in the form of iid pairs

$$(x_1, y_1), \dots, (x_l, y_l), \quad x_i \in X, \quad y_i \in \{-1, +1\} \quad (1)$$

seek a function that approximates the unknown decision rule in the best possible way and provides the smallest probability of incorrect classifications. Training examples are

represented as features x_i and the same feature space is required for predicting future observations. However this approach does not make use of other useful data that is only available at training time; such data is referred to as Privileged Information (PI) (Vapnik and Vashist, 2009). Hence much data that could improve models is set aside and not included in the training process.

In the Learning Using Privileged Information (LUPI) paradigm, training examples instead come in the form of iid triplets

$$(x_1, x_1^*, y_1), \dots, (x_l, x_l^*, y_l), \quad x_i \in X, \quad x_i^* \in X^*, \quad y_i \in \{-1, +1\} \quad (2)$$

where x^* denotes PI. The objective is the same as in classical machine learning, with the extension that privileged information is available in the training stage. One implementation of LUPI is SVM+, and Vapnik and Vashist (2009) showed that this approach and implementation can accelerate the learning process, and outperform classical machine learning in a set of applications.

Conformal prediction (Vovk et al., 2005) is a method that provides a layer on top of an existing machine learning method and uses available data to determine valid prediction regions for new examples. In contrast to standard machine learning that delivers point estimates, conformal prediction yields a prediction region that contains the true value with probability equal to or higher than a predefined level of confidence. Such a prediction region can be obtained under the assumption that the observed data is exchangeable.

In this work we explore conformal prediction in the LUPI paradigm with the aim to improve predictive performance and obtain valid prediction regions. We study the effects of the SVM+ realization of LUPI in an inductive conformal predictor on a benchmark dataset and provide examples in drug discovery problems where machine learning has become a core part of the early discovery process (Norinder et al., 2014; Bendtsen et al., 2017; Zhang et al., 2017). The use of Mondrian conformal predictors is important in drug discovery, where datasets in many cases are imbalanced, and this has been the purpose of several earlier studies (Norinder and Boyer, 2016, 2017; Sun et al., 2017). LUPI opens up for making use of different types of data for the same drug candidate, such as results from different types of experiments. This is currently not used in drug discovery modeling, and there are big opportunities to extend current practices.

2. Data and Methods

2.1. Support Vector Machines (SVM)

Support vector machines (Vapnik, 1998), are one of the most successful methods for classification in machine learning. One of the key concepts of SVM is the use of separating hyperplanes to define decision boundaries, and the optimal decision hyperplane is a plane in a multidimensional space that separates between data points of different classes and that also maximizes the margin, separating the two classes. SVM uses the kernel trick to generate a high dimensional nonlinear representation of the input data examples where it performs the separation with a continuous separation hyperplane, such that the distances of misclassified data examples from the hyperplane are minimized. In this study, we use a classification SVM for training our classification models with a Radial Basis Function

(RBF) kernel

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2),$$

where γ controls the width of the kernel function, and x_i and x_j are the vectors of the i th and the j th training samples, respectively. The kernel parameters γ and the SVM cost parameter C are tuned using two-dimensional cross-validated grid search.

2.2. SVM+

Realizations of LUPI (Vapnik and Vashist, 2009) are mostly based on SVM and referred to as SVM+. In SVM+, the privileged information (PI) is used to estimate the slack variables, which are defined as the distance between the support vectors and the decision boundary. The PI provides a means for regularizing the SVM optimization problem and assists in its generalization. This can be also viewed as augmenting the standard SVM with a second kernel that defines a similarity measure between any two data points in a privileged information space. We use RBF for both kernels, where the first kernel parameters are tuned using SVM on X (standard features) and the second kernel parameter is tuned using SVM on X^* (PI).

2.3. Conformal prediction

Traditional machine learning algorithms for classification problems simply predicts the class labels without any confidence. Conformal predictors expand on this as they output prediction regions for a specific confidence level provided by the user. The confidence value is an indication of how likely each prediction is of being correct, for example, a confidence of 95% implies that the percentage prediction error will be 5% on average. Conformal predictors are built on top of traditional machine learning algorithms, referred to as underlying algorithms, and they can be broadly categorized into transductive and inductive approaches; we refer to Papadopoulos (2008) for more details. We here consider the inductive approach called Inductive Conformal Prediction (ICP), which is more computationally efficient as compared with the transductive approach. In particular, we use Mondrian ICP with SVM or SVM+ as the underlying algorithms, and the SVM or SVM+ distance to the decision boundary to define the non-conformity measures (NCM). Mondrian conformal prediction has the advantage that we achieve validity for the individual classes. To evaluate the performance of conformal predictors, we consider the observed fuzziness, as defined in Vovk et al. (2005).

2.4. Data

As a reference dataset we used the MNIST dataset (LeCun et al., 1998), which has been used previously with the SVM+ algorithm (Vapnik and Vashist, 2009). The MNIST dataset contains grayscale images of handwritten digits 0-9 as vectors of 28 x 28 pixel images, and was downloaded from <http://yann.lecun.com/exdb/mnist/>. We used a 4000 example subset of MNIST dataset comprising digits 5 and 8. The original 28 x 28 pixel images were used as PI, where images resized to 8 x 8 pixel resolution were used as standard dataset. We also used three datasets in drug discovery (Hansen, MMP, and AHR), where chemical

structures are represented as numerical features, and the response variable is measured in a biological assay. The Hansen dataset (Hansen et al., 2009) was constructed to enable the prediction of mutagenicity for the chemical structure of *e.g.* a drug candidate, based on measurements from the Ames Mutagenicity test (Zeiger and Mortelmans, 2001). The MMP dataset is based on measurements for small molecule disruptors of the Mitochondrial Membrane Potential, and is commonly used to assess the effect of chemicals on mitochondrial function (Sakamuru et al., 2016). The AHR dataset is based on measurements for interaction with the aryl hydrocarbon receptor (AHR), related to chemical toxicity and interaction with drugs and other compounds (Bradshaw and Bell, 2009). AHR and MMP were downloaded from PubChem (AHR PubChem AID: 743122, MMP PubChem AID: 720637) as part of the Tox21 project that has previously been used for modeling (Huang et al., 2016). The Hansen dataset was downloaded from <http://doc.ml.tu-berlin.de/toxbenchmark/>. The chemical structures in the datasets AHR, MMP and Hansen were represented using ten Physical-Chemical descriptors (Chi1n, Chi2n, Chi3n, Chi4n, Chi0v, Chi1v, Chi2v, Chi3v, Chi4v and MolLogP), and Morgan fingerprints calculated using RDKit (www.rdkit.org). The Physical-Chemical descriptors contains less features and can be hypothesized to produce less accurate models than Morgan fingerprints. All the datasets are binary class problems with class labels (-1, 1). The details of the datasets are given in Table 1.

Table 1: Description of the datasets and feature sets used in this work

Dataset	Features	# Observations	# Features
MNIST	X: 8×8 pixel images	4000	64
	X*: 28×28 pixel images	4000	784
AHR	X: Phys-chem descriptors	6299	10
	X*: Morgan fingerprints	6299	55725
Hansen	X: Phys-chem descriptors	6509	10
	X*: Morgan fingerprints	6509	48325
MMP	X: Phys-chem descriptors	5647	10
	X*: Morgan fingerprints	5647	49764

2.5. Study design

We denote $X = (x_1, \dots, x_l)^T$, as a matrix of standard features, $X^* = (x_1^*, \dots, x_l^*)^T$, as a matrix of PI, and $y = (y_1, \dots, y_l)$, as a vector of class labels. We chose three statistical models: SVM on X , SVM on X^* and SVM+ on X with X^* as PI. These three models were applied on all four datasets to compare their predictive accuracy and efficiency (observed fuzziness). First, the dataset was partitioned using stratified-split into two parts: training (80%) and external test (20%) set, and the training and the test sets were then kept fixed. Then the training set was randomly divided into proper-training (70%) and a calibration set (30%). For tuning the parameters for each model and for each dataset, we used five-fold cross validation technique on the corresponding proper-training set, and we selected the parameters based on the highest prediction accuracy. More importantly, the tuning was

performed in three steps:

1. Tuning of the first RBF kernel parameter, γ_1 , and the SVM parameter, C_1 , for the model SVM on X : We used two dimensional grid search with 5-fold CV on the X -proper-training set.
2. Tuning of the second RBF kernel parameter, γ_2 , and the SVM parameter, C_2 , for the model SVM on X^* : We used two dimensional grid search with 5-fold CV on the X^* -proper-training set.
3. Tuning of SVM+ parameters, C and γ : We used two dimensional grid search with 5-fold CV on the X -proper-training and X^* -proper-training with selected kernel parameters, γ_1 and γ_2 , in the previous steps.

Table 2: Hyper parameter ranges for various methods

Method	C	γ
SVM on X	[.1, 1000]	[1e-7, 1]
SVM on X^*	[.1, 1000]	[1e-7, 1]
SVM+ on X (with X^* as PI)	[.01, 100]	[1e-4, .1]

The ranges explored for each parameter and for each method are given in Table 2. The proper-training set with corresponding selected parameters was then used to build the model. The non-conformity scores were computed on the corresponding calibration set. We used the SVM/SVM+ decision function to define the non-conformity measure (NCM)

$$\alpha_i = y_i f(x_i),$$

where f is the SVM/SVM+ decision function. Then for each observation in the external test set, we computed (Mondrian) conformal prediction p-values for each class. The above procedure was repeated 10 times, and the average predictive performance, and the average observed fuzziness was reported. The above-mentioned steps are outlined in Algorithm 1.

2.6. Computational Details

The computations were performed on resources provided by SNIC through Uppsala Multi-disciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2017-7-273. We used existing Python implementation of LibSVM in scikit-learn toolkit for training and prediction SVM models. We implemented SVM+ on Python using python-cvxopt: Python package for convex optimization. Conformal prediction using SVM as an underlying machine learning algorithm was implemented in Python using scikit-learn toolkit.

```

Input:  $(X, X^*, y)$ ,  $N$ : number of repetitions
Output: average prediction accuracy, average validity, average observed fuzziness
Step 1: Partition the dataset  $(X, X^*, y)$  into 80% for training, and 20% for testing using
stratified split.
Step 2: Partition the training set,  $(X, X^*, y)$ -training, into 70% proper-training and 30%
calibration.
Step 3: Use cross-validation for tuning  $C_1$  and  $\gamma_1$ , using SVM on  $X$ -proper-training, and
select the one that gives the highest average prediction accuracy.
Step 4: Use cross-validation for tuning  $C_2$  and  $\gamma_2$ , using SVM on  $X^*$ -proper-training, and
select the one that gives the highest average prediction accuracy.
Step 5: Use cross-validation for tuning  $C$  and  $\gamma$ , using SVM+ on  $X$ -proper-training with
 $X^*$ -proper-training as PI, and the kernel parameters  $\gamma_1$  and  $\gamma_2$  as selected in Step 3 and
Step 4 respectively.
Step 6:
repeat
  Step 6.1: Randomly partition the  $(X, X^*, y)$ -training set into 70% for proper-training,
  and 30% for calibration.
  Step 6.2: Train the three models using their corresponding proper-training set.
  Step 6.3: Compute NCM using the corresponding calibration set for each model.
  Then for each model compute the following on their corresponding test set:
  - prediction accuracy
  - deviation from exact validity
  - observed fuzziness
until  $N$  iterations;
Return average prediction accuracy, average validity and average observed fuzziness of each
model.

```

Algorithm 1: Algorithm for study design: ICP with SVM and SVM+

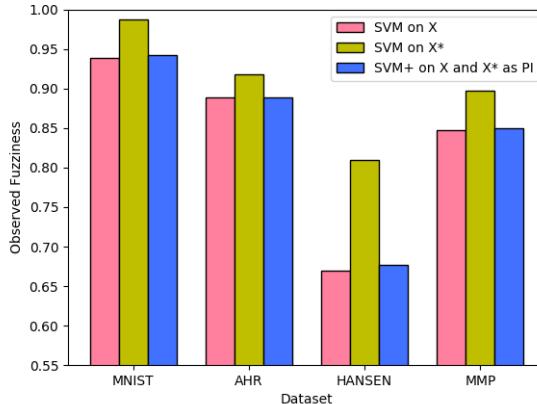
3. Results and Discussion

In this study, we have used prediction accuracy and observed fuzziness as measures of performance. The prediction accuracies of the three statistical models are given in Table 3 and in Figure 1, and we observe that the methodology using SVM on X^* outperforms the other models in terms of prediction accuracy for all datasets, but we note that SVM+ outperforms SVM on X for most of the datasets.

For comparison of the three Mondrian inductive conformal predictors, their measure of efficiency and validity are given in Table 4 and in Figure 2. The smaller the efficiency (observed fuzziness) is, the better the model performs. Also here we see that SVM on X^* performs best in terms of efficiency, and that SVM on X^* outperforms SVM on X for all datasets, but that the level of improved efficiency with SVM+ varies between the datasets. One implication of using SVM+ is the need for tuning additional SVM hyper-parameters associated with PI, which increases the computational complexity substantially.

Table 3: Comparison of prediction accuracy

Dataset	Statistical Model	prediction accuracy
MNIST	SVM on X	0.939125
	SVM on X^*	0.987375
	SVM+ on X with X^* as PI	0.942875
AHR	SVM on X	0.888889
	SVM on X^*	0.917857
	SVM+ on X with X^* as PI	0.888889
Hansen	SVM on X	0.669124
	SVM on X^*	0.809370
	SVM+ on X with X^* as PI	0.676651
MMP	SVM on X	0.847522
	SVM on X^*	0.896726
	SVM+ on X with X^* as PI	0.849292

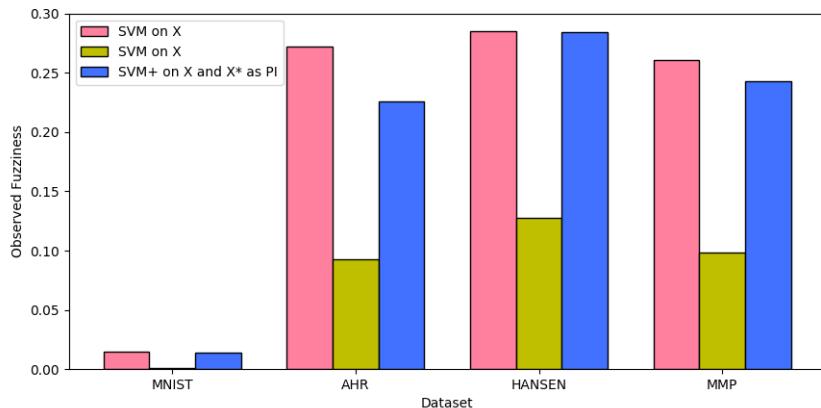
Figure 1: Comparison of prediction accuracy on four selected datasets using SVM on X (pink), SVM on X^* (yellow) and SVM+ (blue).

4. Conclusions

We here introduced conformal prediction using LUPI/SVM+ as underlying method. We investigated the validity and efficiency of inductive conformal predictors with SVM+ on the MNIST benchmark dataset, and also applied it to three datasets relevant to drug discovery. Our results show that the ICP with SVM+ is more efficient than ICP with SVM on X, in terms of observed fuzziness. We also showed that the prediction accuracy of SVM+ on X with X^* as privileged information is better than standard SVM on X for all datasets, however in some cases the improvements on observed fuzziness and prediction accuracy are only marginal. With confidence in predictions being central for drug discovery decision

Table 4: Comparision of validity and efficiency

Dataset	Learning Algorithm	Validity	observed fuzziness
MNIST	SVM on X	0.189254	0.015103
	SVM on X^*	0.182684	0.000839
	SVM+ on X with X^* as PI	0.176197	0.013733
AHR	SVM on X	0.168761	0.272146
	SVM on X^*	0.107204	0.092754
	SVM+ on X with X^* as PI	0.100159	0.226047
Hansen	SVM on X	0.121286	0.285467
	SVM on X^*	0.128802	0.127245
	SVM+ on X with X^* as PI	0.130737	0.283943
MMP	SVM on X	0.164847	0.260612
	SVM on X^*	0.140734	0.098628
	SVM+ on X with X^* as PI	0.151952	0.248843

Figure 2: Comparision of observed fuzziness on four selected datasets using SVM on X (pink), SVM on X^* (yellow) and SVM+ (blue).

making, LUPI in a conformal prediction setting has high potential to become a useful methodology in pharmaceutical development.

Acknowledgments

We would like to acknowledge Alexander Kensert and Jonathan Alvarsson for assistance in data preparation. This project received financial support from the Swedish Foundation for Strategic Research (SSF) as part of the HASTE project under the call 'Big Data and Computational Science'. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project SNIC 2017/7-273. LC is grateful for the support given by the ExCAPE project funded by the European Unions Horizon 2020 Research and Innovation programme under Grant Agreement no. 671555 and by the Swedish Knowledge Foundation through the project Data Analytics for Research and Development (20150185).

References

Claus Bendtsen, Andrea Degasperi, Ernst Ahlberg, and Lars Carlsson. Improving machine learning in early drug discovery. *Annals of Mathematics and Artificial Intelligence*, 81(1-2):155–166, 2017.

Tracey D Bradshaw and David R Bell. Relevance of the aryl hydrocarbon receptor (ahr) for clinical toxicology. *Clin Toxicol (Phila)*, 47(7):632–42, Aug 2009. doi: 10.1080/15563650903140423.

Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of Chemical Information and Modeling*, 2009. doi: 10.1021/ci900161g. URL <http://dx.doi.org/10.1021/ci900161g>.

Ruili Huang, Menghang Xia, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Matthias Attene-Ramos, Tongan Zhao, Christopher P Austin, and Anton Simeonov. Modelling the tox21 10 k chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat Commun*, 7:10425, Jan 2016. doi: 10.1038/ncomms10425.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ulf Norinder and Scott Boyer. Conformal prediction classification of a large data set of environmental chemicals from toxcast and tox21 estrogen receptor assays. *Chem Res Toxicol*, 29(6):1003–10, Jun 2016. doi: 10.1021/acs.chemrestox.6b00037.

Ulf Norinder and Scott Boyer. Binary classification of imbalanced datasets using conformal prediction. *J Mol Graph Model*, 72:256–265, Mar 2017. doi: 10.1016/j.jmgm.2017.01.008.

Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *J Chem Inf Model*, 54(6):1596–603, Jun 2014. doi: 10.1021/ci5001168.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. InTech, 2008.

Srilatha Sakamuru, Matias S Attene-Ramos, and Menghang Xia. Mitochondrial membrane potential assay. *Methods Mol Biol*, 1473:17–22, 2016. doi: 10.1007/978-1-4939-6346-1_2.

Jiangming Sun, Lars Carlsson, Ernst Ahlberg, Ulf Norinder, Ola Engkvist, and Hongming Chen. Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. *J Chem Inf Model*, 57(7):1591–1598, 07 2017. doi: 10.1021/acs.jcim.7b00159.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: learning using privileged information. *Neural Netw*, 22(5-6):544–57, 2009. doi: 10.1016/j.neunet.2009.06.042.

Vladimir N Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications and control series. John Wiley & Sons, New York. A Wiley-Interscience Publication, 1998.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

E Zeiger and K Mortelmans. The salmonella (ames) test for mutagenicity. *Curr Protoc Toxicol*, Chapter 3:Unit3.1, May 2001. doi: 10.1002/0471140856.tx0301s00.

Lu Zhang, Jianjun Tan, Dan Han, and Hao Zhu. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today*, 2017.