Journal of Machine Learning Research 1–16

Technical report (DRAFT) on multi-target learning

Ilia Nouretdinov, Khuong Nguyen, Alex Gammerman
Computer Learning Research Centre
Royal Holloway, University of London, London, UK

Abstract

This report describes the work in progress, analysing ExCAPE data on possibility of multi-target learning. We start with observing the structure of missing values (labels), as the sets of examples overlap but are not identical for different targets. Then we concentrate on the part of the data with full information in order to consider mutual dependence between the targets, and possibility of improvement of prediction by collecting the information together.

1. Introduction

Most of the work, done by CLRC group for ExCAPE project before (such as (1; 2)), was devoted to making reliable predictions for different targets independently of each other. Each of them was considered as a separate binary classification problem, answering the question whether a compound is active on a specific target. One of the causes for that was incomplete correspondence between the information available for different targets. For a large part compounds, their activity was measured only on some of the targets, not all of them.

However, it may happen that information about each target gives a 'hint' for prediction of other targets. Here we try to cover this gap.

For this preliminary study, we have taken three top targets, for which the largest amount of information is available. As we will see, the activities are positively dependent on each other, however the structure of missing values is also not random. However, on some clause it is possible to restrict us only to the examples where information about all three targets is available, and assume that the training an testing set are shared for all the three problems.

Joint classification for three targets, based on a shared training set can have interpretation in terms of multi-target learning (5) or LUPI paradigm (**?**). Detailed review of them, with their combination with conformal framework, can be found in .

However, for a concrete data base this is a challenge to estimate whether any of this methods can increase the efficiency of the prediction. The key question can be posed in this way: does learning on two or more targets improve the prediction for one of the targets, compared to the prediction to the prediction after training without involving information for the other targets? We try to show that the answer is positive for ExCAPE data, and by the way formulate a modified algorithm scheme for approaching LUPI task, inspired by this data analysis.

2. Studies on data label structure

The information for each data example (compound) in ExCAPE project originally consists of its sparse (QSAR) feature vector, and its activity on some number of targets.

Here we start with analysing data for possible conclusions the structure of the information for data labels, without going into feature vectors for further sections. This study was done on three targets.

2.1. Data for three targets

For the initial analysis we have taken three top targets:

1. CFTR;
2. IDH1;
3. NFE2L2.

Originally, activity is a numerical value, but we prefer to restrict us to a binary classification problem now, not regression. Therefore, each of them divides the data into three categories:

1. not active (activity<5);
2. active (activity>5);
3. unknown (not presented).

Table 1 shown the generals statistic of labels. By examples with unknown labels for one of the targets we mean the case when the labels is known for one or two other targets. The example for which all three labels are unknown, are left out of consideration.

Table 1: Data distribution

	inactive	unknown	active
CFTR	458748	67767	1166
IDH1	462707	58883	6091
NFE2L2	386280	113171	18130

Table 2 shows pairwise joint distributions that would allow further to study dependence/correlation between the activities themselves, and between the activities and availability of knowledge.

The overall (three-target) joint distribution is shown in Table 3. The 'middle' cell contains 0 by the reason mentioned above. This table is shown for information and does not play important role in the study.

Table 2: Data distribution (pairwise)

(any NFE2L2)	IDH1 inactive	IDH1 unknown	IDH1 active
CFTR inactive	406642	47024	5082
CFTR unknown	55341	11703	723
CFTR active	724	156	286
(any IDH1)	CFTR inactive	CFTR unknown	CFTR active
NFE2L2 inactive	336765	48663	852
NFE2L2 unknown	98442	14590	139
NFE2L2 active	23541	4514	175
(any CFTR)	IDH1 inactive	IDH1 unknown	IDH1 active
NFE2L2 inactive	370715	11251	4314
NFE2L2 unknown	66976	45886	309
NFE2L2 active	25016	1746	1468

Table 3: Data distribution (triplewise)

NFE2L2 inactive	IDH1 inactive	IDH1 unknown	IDH1 active
CFTR inactive	331753	1192	3820
CFTR unknown	38355	10034	274
CFTR active	607	25	220
NFE2L2 unknown	IDH1 inactive	IDH1 unknown	IDH1 active
CFTR inactive	52676	45757	9
CFTR unknown	14292	0	298
CFTR active	8	129	2
NFE2L2 active	IDH1 inactive	IDH1 unknown	IDH1 active
CFTR inactive	22213	75	1253
CFTR unknown	2694	1669	151
CFTR active	109	2	64

2.2. Missing labels

Observing Table 2, we can state the following question for each pair (A, B) of different targets. Compare the examples (A) for which the activity of target A is known, and $(A*)$ where it unknown, and check whether the distribution of activity of a target B is different for these two groups.

The problem is well-known as so called Missing Not At Random (MNAR) data. "You can then run t-tests and chi-square tests between this variable and other variables in the data set to see if the missingness on this variable is related to the values of other variables"(3).

To test this, χ^2-test was applied to the contingency table according to the methodology shown in Table 4. If $ad > bc$ significantly, this means that information on target A tends to be missing if target B is active, the opposite is if $ad < bc$ significantly.

Table 4: Contingency table for targets A,B

	B: inactive	B: active	B: unknown
A: active or inactive	a	b	-
A: unknown	c	d	-

The results of the χ^2-test are shown in Table 5, where some kind of significant MNAR effect is found for 5 of 6 target pairs having sense in this context. The direction of dependence is selected by comparison of ad and bc as mentioned above.

Table 5: The results of χ^2 test

Target A (known/unk.)	Target B (act./inact.)	p-value (chi-square)	direction of dependence
CFTR	IDH1	0.83	–
CFTR	NFE2L2	7.9×10^{-61}	CFTR unknown∼active NFE2L2
IDH1	CFTR	7.2×10^{-28}	IDH1 unknown∼active CFTR
IDH1	NFE2L2	1.0×10^{-203}	IDH1 unknown∼active NFE2L2
NFE2L2	CFTR	2.2×10^{-15}	NFE2L2 unknown∼inactive CFTR
NFE2L2	IDH1	5.1×10^{-96}	NFE2L2 unknown∼inactive IDH1

2.3. Is the data collection biased?

Multiple low p-values is Table 5 signify possible bias in data collection. Its most likely explanation is that active examples are under-represented in data collected for CFTR and IDH1, and/or over-represented in NFE2L2.

This may be taken into account in the interpretation of predictions. However, such explanation is incomplete: for example, there may be only under-representation in CFTR/IDH1, or only over-representation in NFE2L2, and these two causes are not distinguishable by means of data analysis only.

According to the nature of problem, we can assume that latter case (over-representation) is more likely, because typically positive examples are more of interest for data collection, registered more frequently, and therefore may be over-represented. In that case, p-values and probabilistic scores assigned to prediction of activity the compounds by machine learning algorithms may be slightly shifted to upper side.

However, we see that the degree of over-representation depends on the specific kind of target. In this case the degree is the largest of three in NFRE2L2 that is also the smallest by overall amount of collected information, but the largest in the number of registered actives (see Tab.1). It looks like that the collection was more accurate for two top targets (CFTR and IDH1) and more restricted to positives for this one. Checking more targets may clarify how typical this effect is.

2.4. Data with complete information

Up to clauses made above about possible shifts, we will restrict us to the example where information about activity is present for all three targets.

The extraction from the data summary is shown in Table 6. It shows three-target joint distribution of activity, as well as its two- and one-dimensional summaries for target pairs and unique targets. 'Any IDH1' etc. in this table refer only to the cases when the values are known, not missing.

Table 6: Distribution of data without missing labels

NFE2L2 inactive	IDH1 inactive	IDH1 active	(any IDH1)
CFTR inactive	331753	3820	335573
CFTR active	607	220	827
(any CFTR)	332360	4040	336400
NFE2L2 active	IDH1 inactive	IDH1 active	(any IDH1)
CFTR inactive	22213	1253	23466
CFTR active	109	64	173
(any CFTR)	22322	1317	23639
(any NFE2L2)	IDH1 inactive	IDH1 active	(any IDH1)
CFTR inactive	353966	5073	359039
CFTR active	716	284	1000
(any CFTR)	354682	5357	360039

Now we apply χ^2-test just to study dependence between two kinds of activity. The methodology is shown in Tab. 7.

Table 7: Contingency table for targets A,B

	B: inactive	B: active	B: unknown
A: active	a	b	-
A: inactive	c	d	-
A: unkwown	-	-	-

The results, presented inTable 8, show that it is significant for all 3 pairs, and the correlation is positive. This actually means that knowledge of one of them may help in prediction of the other.

Table 8: Results of χ^2 check

Target A	Target B	p-value	direction
CFTR	IDH1	$< 10^{-50}$	CFTR active∼active IDH1
CFTR	NFE2L2	7.2×10^{-40}	CFTR active∼active NFE2L2
IDH1	NFE2L2	$< 10^{-50}$	IDH1 active∼active NFE2L2

3. Multi-target data distibution

The previous section was finished on a conclusion about high dependence between data activities. This definitely should help to predict for a compound its activity on one target based on information for the other targets. But what we here wish so state a less obvious question: whether the prediction quality of one target for a compound can be improved if the information about other targets for the training compounds, not for same one? This is actual in assumption that we are working with a new compound for which nothing is known about its activity.

To answer this question, some study of feature vectors is also needed. In this section we involve feature vectors into the analysis. The aim is to make conclusions about possibility of joint prediction of labels for different targets, as something more informative compared to their independent prediction.

3.1. Distribution of feature vectors

Leet us start with a multi-target visualisation. Fig.1 contains combined plot of the data vectors with respect to three types of activities (see details in the caption), where the dimensionality reduction was done by well-known distance-based tSNE approach. Only 1% of the data is presented on the plot, therefore only some of possible activity combinations are presented.

Let us use notation [A+]/[A-] for the class of examples with positive/negative label for target A, and [A+B-] for intersection of two such classes for different targets A and B. Observing Fig.1, we seems that combined [CFTR+IDH1+] activity is the most compactly located (and therefore, potentially predictable) combination. It looks much more recognisable than each of CFTR+ and IDH1+ as they are.

This leads to an idea to consider the 'combined' prediction of [CFTR+IDH1+] as a separate auxiliary task. The working hypothesis is its usability to improve the prediction quality of CFTR and/or IDH1.

3.2. Approach for joint prediction

The next stage of data analysis is based on modelling a prediction algorithm that incorporates the additional information coming from 'combined' prediction. By 'combined' prediction we mean the straightforward prediction of [CFTR+IDH1+] as a logical conjunction: the 'combined' label is positive if it is positive for both of the combined targets.

We consider interpretation of the prediction task as a kind of LUPI (6) problem, as fat as we do not intend to use labels of the testing examples, and the goal for evaluation is prediction quality of one of the targets.

Like (9), we use elements of Inductive Conformal Prediction (a calibration set), but the suggested scheme is different from that work.

We assume that there is a natural underlying binary classification method that assigns a testing example numerical scores in favour of the hypothesis of its positive activity, and it is naturally adoptable to the 'combined' prediction that is also a binary one.

The plan is summarised as follows.

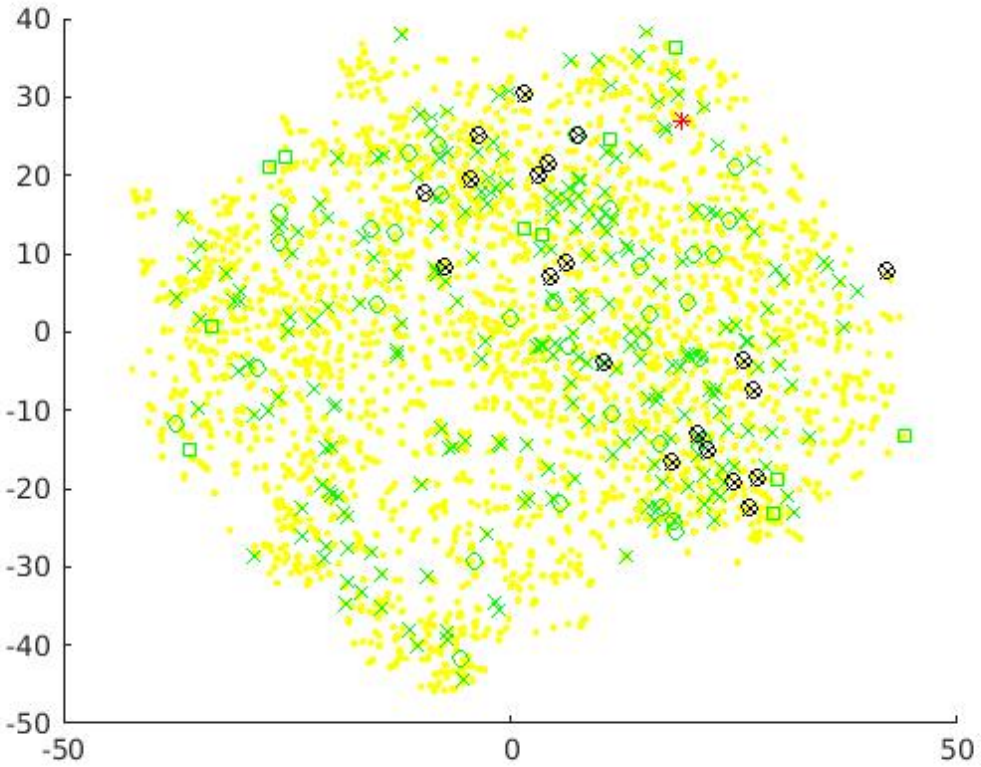

Figure 1: Yellow: none of activities; Green: 1 activity; Black: 2 activities; Red: 3 activities; Shape: kind of activity (CFT=square, IDH1=circle, NFE2l2=cross). Shown for randomly selected 1% of the data in tSNE projection.

1. Calibration. As in the inductive mode of prediction, we divide original training data set into the proper training set (further referred as training), and the calibration set. For both of these set we assume the information about both kinds of activity to be available, unlike the testing set.

2. Scoring. After learning on the training set, two types of scores (standard and 'combined') are provided for each of calibration/testing examples. Thus, the original calibration/testing feature vectors are converted for them to two-dimensional score vectors, that are then used for label predictions.

3. Re-training. It is done on the calibration set in order to obtain the prediction rule for converting the score vectors into labels.

4. Testing. The prediction rule is applied to testing set in order to make new predictions.

3.3. Experimental details

We use the following random data split: 200,000 training examples, 100,000 calibration examples, 112,724 testing examples, and consider the combination of two tasks: prediction of activity for target 1 (either CFTR or IDH1), and prediction of the combined activity [CFTR+IDH1+], also interpreted as a binary classification task.

As the underlying method for initial numerical scores of both standard and combined tasks, we use k-NN (k nearest-neighbours) method, selected for its best performance on the standard task. As our scores, we take k-NN scores calculated as: 'average distance to k nearest inactive examples, divided by average distance to k nearest active examples' where neighbours are taken within the proper training set. In the context of conformal prediction, they were used as conformity scores for the hypothesis 1, or non-conformity scores for the hypothesis 0. However, here we do not apply the conformal prediction yet, and use the scores in another way.

The parameter is set to $k = 10$ due to its best performance for the standard activity prediction tasks.

Two alternatives methods (SVM and xGB) were tried but left out, as they perform well on the standard task but fail on the highly imbalanced combined [CFTR+IDH1+] task, producing predictions that even do not show any significance dependence on the true values. This at least means that tuning of k-NN method is more transferrable, that is another reason we concentrate on this method for initial modelling.

Re-training will also be done by K-nearest neighbours method, but with much larger number of neighbours than on the step of score vector creation, that allow to simplify the way of further prediction by using 'voting' approach: 'how many of K neighbours do have the label 1'.

3.4. Visualisation of score vectors

Let us go from feature vectors to score vectors.

Fig. 2 shows the empirical joint distributions of k-NN scores for non-activity hypothesis on the calibration/testing set. Remind that the large value of the score corresponds to high evidence in favour of the activity hypothesis, against non-activity.

The first axis on Fig. 2 corresponds to prediction of one of the activities (CFTR on the left, IDH1 on the right) without any usage of the other. A natural simple way of standard prediction is using a threshold that would correspond to a vertical line on this plot.

The second axis is the same kind of two-class non-conformity score applied to the join prediction problem, when only the 'combined' activity (overlap of activity on two targets) is considered as the positive class, and all the rest is considered negative. By visual inspection of the plots, it can be formulated hypothetically that a positive prediction is more trustable if its combined score is close to the standard scores. However, there is no evident way of division between 'blue' and 'yellow' areas on this plot, therefore we rely on K-NN approach on this step.

The colour of the examples corresponds to positive class of the activity being predicted (target 1 is either CFTR or IDH1).

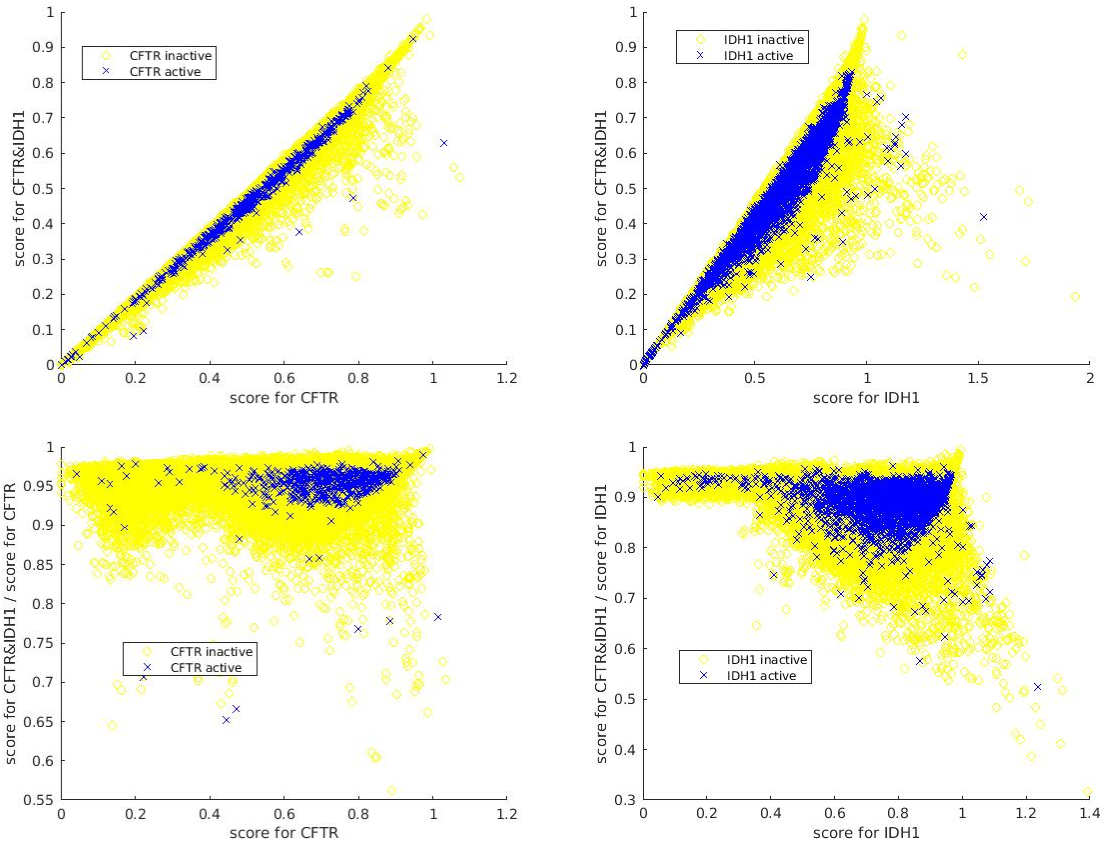

Figure 2: k-NN ($k = 10$) scores for single activity with support of combination scores

3.5. Prediction curves

The final goal is comparison of the results in terms of prediction curves.

The result is shown on Fig. 3. It contains the following curve

- Random baseline (red): prediction quality that may be achieved without any training by random prediction.
- No additional information (green): using only standard scores.
- Combined prediction (black): we show the results which were achieved with parameters $k = 10$ (for initial scores); $K = 10,000$ (for re-training on score vectors).

The prediction curves are parametrised with variable threshold t for the decision rule.

For the baseline (ignoring additional information) possible decision rules are 'output a positive prediction for a testing example if its score for target 1 is at least t'. For the proper prediction it is: 'output a positive prediction for a testing example if the proportion of positives amongst K closest calibration score vector is at least t'.

Tab. 9 shows areas under curve for $k = 10$ and different values of the parameter K.

Table 9: Areas Under Curve

k	K	AUC (CTFR)	AUC (IDH1)
random baseline		0.5	0.5
10	no add. info	0.5415	0.5690
10	100	0.5090	0.5452
10	1,000	0.5273	0.5648
10	10,000	0.5664	0.5523

Table 10: Negative class accuracy for positive class accuracy 0.9

k	K	$\mathrm{ACC}_{0.9}$ (CTFR)	$\mathrm{ACC}_{0.9}$ (IDH1)
random baseline		0.1	0.1
10	no add. info	0.1865	0.1875
10	1,000	0.1148	0.1891
10	10,000	0.1776	0.1913

Observing the results, we see obvious improvement for CFTR. Additional information coming from combined prediction improves the whole curve.

The result of the prediction of comparison for IDH1 is not successful in general, but some improvement is present if the required accuracy of active example is close to 1. It shows an advantage if we evaluate the results by the following criterion: 'what accuracy is achievable on the negative class if the accuracy on positive class is 0.9?' (here 0.9 is taken as an example: assumed that the prior requirement is to catch at least 90% active examples).

This is shown in Tab. 10 and illustrated by Fig. 4 for prediction of IDH1, showing the corresponding part of the plot.

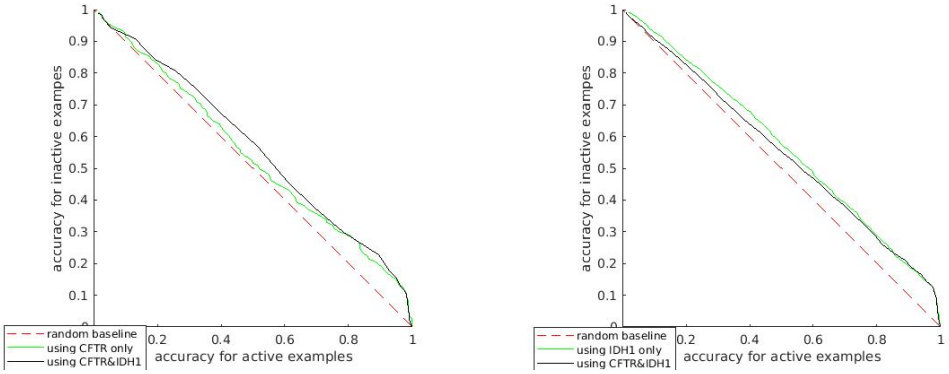

Figure 3: Prediction curves compared ($k = 10$, $K = 10,000$)

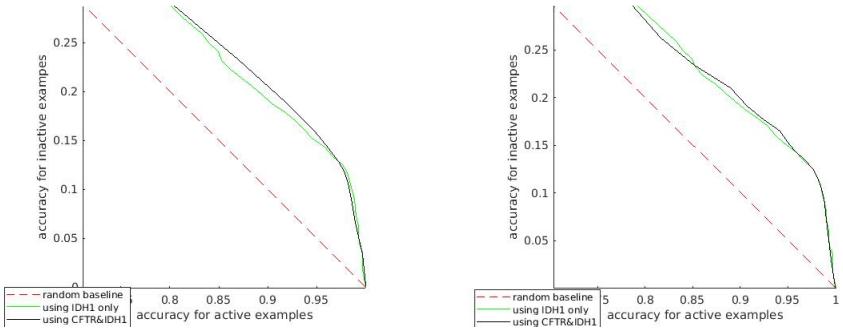

Figure 4: IDH1 prediction curves compared ($k = 10$; $K = 1000$ and $K = 10,000$)

4. Conclusions and plans of the future work

According to the analysis, provided in this paper, we state several possible directions for conclusions and next steps.

4.1. Missing labels

In Sec. 2.2 we have detected a possible connection bias that may be taken into account in the interpretation of predictions: for some of the targets the predictions are more likely to contain a positive shift, while for the others it may be negative. This might be analysed on more targets.

4.2. Joint prediction

The multi-class analysis for the data with full information about labels leads to the following recommendation: to complement the prediction of already implemented $[A+]$ and $[B+]$ bt extra prediction of $[A + B+]$. Observing Fig. 2 leads to a hypothesis that the privileged information may be used in its 'support' role: when one of the targets is being predicted, a positive prediction is more trustable if it is supported by prediction made after training only on the 'overlapping' examples (with more than one activity) as positives. This also may have such interpretation as 'overlapping' examples being marked by an expert as the most 'strong' and trustable positive examples.

Preliminary investigation (Sec.3.2) also inspires a possible novel approach for LUPI prediction: to make 'normal' prediction based on the actual data, and to support it with additional prediction where the active class is replaced with the smaller one containing 'overlapping' examples only. Some calibration set can be used to find the best way of including the information coming from the additional prediction.

An advantage of this approach is that it provides a visual way of preliminary assessment of the data on 'LUPI/joint prediction'. We try to obtain an evidence in favour or against the possibility of effective usage of additional information (that in the considered case comes from other targets) for improvement of classification quality. Surely, the potential usage of 'Scoring' step is not limited to 'combined' additional scores calculated exactly as in the provided example; this is generalisable to other ways of processing combined info.

However, in this report we were less concerned with developing concrete ways of reliable prediction. The task for the future is integration of the suggested scheme with conformal or Venn prediction. The easiest way is to do it on the stage of re-training, using voting results of K-NN method to created non-conformity scores.

4.3. Partial LUPI paradigm

Another actual understanding of the problem was suggested by P.Toccacelli after discussing this work. LUPI paradigm can be extended to the case when the privileged information is present only in part of the data. This would allow to train on all the existing example, not restricting us to ones with full information. In some sense, this paradigm was earlier used in (10), although the implementation was done in the context of Venn machines.

Acknowledgements

This work was supported by European Union Grant 671555 ("ExCAPE"), AstraZeneca grant "Machine Learning for Chemical Synthesis" (R10911).

References

[1] Toccaceli, P., Nouretdinov, I., Gammerman, A. Conformal Predictors for Compound Activity Prediction 17 Apr 2016 Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016 Madrid, Spain, April 20?22, 2016 Proceedings. Springer, Vol. 9653, p. 51-66 (Lecture Notes in Computer Science; vol. 9653)

[2] Toccaceli, P., Nouretdinov, I., Gammerman, A. Prediction of Biological Activity of Chemical Compounds 16 Jun 2017 In : Annals of Mathematics and Artificial Intelligence. p. 1-19

[3] Grace-Martin, K. How to Diagnose the Missing Data Mechanism. On-line review: http://www.theanalysisfactor.com/missing-data-mechanism/

[4] Mair, Patrick, Kurt Hornik, and Jan de Leeuw. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. Journal of statistical software 32.5 (2009): 1-24.

[5] Wang, H., Liu, X., Nouretdinov, I., Luo, Z. A Comparison of Three Implementations of Multi-Label Conformal Prediction Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings. pp.251–259

[6] Vapnik, V., Izmailov, R. Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer. Statistical Learning and Data Sciences. Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings, pp.3–32.

[7] Yang, M., Nouretdinov, I., Luo, Z. Learning by Conformal Predictors with Additional Information The 9th Artificial Intelligence Applications and Innovations Conference (AIAI): 2nd Workshop on Conformal Prediction and its Applications, (IFIP Advances in Information and Communication Technology; vol. 412), 2013, pp. 394–400.

[8] Nouretdinov, I. Improving reliable probabilistic prediction by using additional knowledge. 6th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2017).

[9] Gauraha, N., Carlsson, L., Spjuth, O. Conformal Prediction in Learning Under Privileged Information Paradigm with Applications in Drug Discovery To appear in: 7th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2018).

[10] Nouretdinov, I. Improving reliable probabilistic prediction by using additional knowledge. 6th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2017). p. 193-200

[11] Vovk, V., Fedorova, V., Nouretdinov, I., Gammerman, A. Criteria of efficiency for conformal prediction. Proceedings of the 5th International Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2016).

Appendix: Literature Review on Conformal Prediction in LUPI

ExCAPE project includes prediction of activity for different targets for the same compound. Currently, the prediction for each target is considered as a separate machine learning problem, being solved independently of the others.

Observation of the relevant literature have shown that this may lead to some loss of efficiency for machine learning. So it worth to try the following alternatives.

First, multi-label prediction suggested in (5). In the context of ExCAPE project, multi-label means the same as multi-target prediction.

Second, interpretation in terms of Learning Under Privileged Information (6): when activity for one target is predicted, activity for the other targets is understood as additional (privileged) info.

Let us state the problem and possible solution formally. Each compound x_i is assigned a feature vector $(x_i^1, \ldots, x_i^m)$ and a label vector $(y_i^1, \ldots, y_i^q)$, where m is the number of features and q is the number of targets.

The compounds $x_1, \ldots, x_n$ belong to the training set (in inductive mode, it is split into proper training and calibration parts), x_{new} is one of testing compounds, with known feature vector $(x_{new}^1, \ldots, x_{new}^m)$.

The current (basic) approach is Alg. 1. It can be used as a baseline for other methods. Its possible alternatives are: multi-label prediction (5) (Alg. 2), and LUPI prediction (6) (Alg. 3).

Algorithm 1 Basic (lower baseline) prediction scheme

FOR $j = 1, \ldots, q$
train on $(x_i^1, \ldots, x_i^m; y_i^j)$ where $i = 1, \ldots, m$
predict y_{new}^j for $(x_{new}^1, \ldots, x_{new}^m)$
END FOR

Multi-label prediction

The scheme of multi-label prediction (5) is given by Alg. 2. Original meaning of the term is making prediction in assumption that different labels do not exclude each other. It is also possible that none of the labels is actual ('empty' case) although it was believed to be a rare event. Therefore, for q labels, there are 2^q possible answers. The same number of answers appear in the multi-target problem, if we consider q targets as analogue of q non-exclusive labels. But it has to be taken into account that the 'empty' case is not negligible: it is expected to appear much more frequently than in the examples from (5). This leads to necessity of some modifications.

Algorithm 2 Multi-label prediction scheme

train on $(x_i^1, \ldots, x_i^m; y_i)$ where $i = 1, \ldots, m$, $y_i = (y_i^1, \ldots, y_i^q)$
predict $y_{new} = (y_{new}^1, \ldots, y_{new}^q)$ for $(x_{new}^1, \ldots, x_{new}^m)$

The work (5) describes several ways of its realisation within the conformal framework, with prediction region as the output. In terms of the ExCAPE multi-target problem described above, they need to be re-interpreted and modified as follows.

1. Power Set MLCP: to test each of 2^q possible hypotheses about y_{new} within multi-class conformal framework, output the set of combinations accepted at level ε. This method need accurate formulating of appropriate criterion of efficiency (will be discussed below), and accurate selection of NCM for 2^q class problem that satisfies it in the best way.

2. Binary Relevance MLCP: a practical analogue of the baseline approach (Alg. 1).

3. Instance Reproduction MLCP: to formulate a new $(q+1)$-class problem and to create a new training set. If y_i is a vector of zeros, x_i is included with label 0. If the vector y_i contains 1s, then x_i is included as many times as many 1s there are, each time with another label j such that $y_i^q = 1$. For example, if $q = 5$ and $y = (0, 1, 0, 0, 1)$, then the example is included twice, with labels 2 and 5. For a new example x_{new}, each of $q+1$ hypotheses is tested, and the prediction region consists ones accepted at level ε. Strictly saying, validity of this method is not quite rigorous, but it is shown in (5) to be practically plausible.

Learning under privileged information

Multi-target prediction may be also understood in terms of LUPI (6), the scheme is given by Alg. 3. LUPI may be more convenient in comparability with the basic approach than multi-label prediction. There is no need for special efficiency criteria as in (5), the general conformal efficiency criteria suggested in (11) are applicable.

Before applying LUPI, it may have sense to try the upper baseline (Alg. 4). If it does not show an improvement compared to Alg. 3 then there is no much reason to expect it from LUPI as well, because the efficiency of LUPI application (Alg. 3) is naturally expected to lie between Alg. 1 as the lower baseline and Alg. 4 as the upper baseline.

The works (7; 8) include possible realisations of LUPI within conformal and Venn frameworks. The approach suggested in (7) appears to be similar to Power Set MCLP.

Algorithm 3 LUPI prediction scheme

FOR $j = 1, \ldots, q$
train on $(x_i^1, \ldots, x_i^m; x_i^*; y_i^j)$ where $i = 1, \ldots, m$, $x_i^* = (y_i^1, \ldots, y_i^{j-1}, y_i^{j+1}, \ldots, y_i^q)$
predict y_{new}^j for $(x_{new}^1, \ldots, x_{new}^m)$
END FOR

Algorithm 4 Upper baseline for LUPI prediction scheme

FOR $j = 1, \ldots, q$
train on $(x_i^1, \ldots, x_i^m; x_i^*; y_i^j)$ where $i = 1, \ldots, m$, $x_i^* = (y_i^1, \ldots, y_i^{j-1}, y_i^{j+1}, \ldots, y_i^q)$
predict y_{new}^j for $(x_{new}^1, \ldots, x_{new}^m; x_{new}^*)$
END FOR