Towards Better Evaluation of Multi-Target Regression Models *

Evgeniya Korneva[1] and Hendrik Blockeel[1,2]

[1] KU Leuven, Department of Computer Science
Celesteijnenlaan 200A, 3001 Leuven, Belgium
[2] Leuven AI
firstname.lastname@kuleuven.be

Abstract. Multi-target models are machine learning models that simultaneously predict several target attributes. Due to a high number of real-world applications, the field of multi-target prediction is actively developing. With the growing number of multi-target techniques, there is a need for comparing them among each other. However, while established procedures exist for comparing conventional, single-target models, little research has been done on making such comparisons in the presence of multiple targets. In this paper, we highlight the challenges of evaluating multi-target models, focusing on multi-target regression algorithms. This paper reviews the common practice and discusses its shortcomings, indicating directions for future research.

Keywords: multi-task learning · multi-target regression · evaluation

1 Introduction

Multi-target learning refers to building machine learning models that are capable of simultaneously predicting several target attributes, which allows the model to capture inter-dependencies between the targets and, as a result, make better predictions. If the target attributes are binary, the problem is referred to as multi-label classification. Multi-dimensional classification is a more general setting where each instance is associated with a set of non-binary labels. Multi-target regression problems, in turn, refer to predicting multiple numerical attributes at the same time.

Due to a large number of real-world applications, the field of multi-target prediction is rapidly expanding. Multi-target problems often occur in ecological modelling, bioinformatics, life sciences, e-commerce, finance, etc. Consider, for instance, predicting several water or air quality indicators (multi-target regression) or product or text categorization (multi-label classification).

Many widely-used machine learning algorithms have been extended towards multi-target prediction. In addition, various specialized methods have been designed to tackle multi-target prediction tasks.

* This research is supported by Research Foundation - Flanders (project G079416N, MERCS)

Approach	Method	References	Year
Problem transformation	Random linear target combinations	Tsoumakas et al. [17]	2012
	Regressor chains	Spyromitros-Xioufis et al. [15]	2016
	SVR	Melki et al. [13]	2017
Algorithm adaptation	Multi-target regression trees	Kocev et al. [11]	2013
		Breskvar et al. [3]	2018
	Rules	Aho et al. [1]	2012
	Low-rannk learning	Zhen et al.[21]	2017
	Neural networks	Hadavandi et al. [8]	2015
	Multi-target SVR	Xu et al. [19]	2013
		Tuia et al.[18]	2017

Table 1: To review the common practice in evaluation of multi-target regression models, we consider ten representative papers from the field.

The more algorithms are being proposed to solve multi-target problems, the higher the need to compare them among each other is. However, no methodology to properly evaluate multi-target algorithms has been developed so far. There exist established techniques for comparing conventional, single-target models, but they are not directly applicable in the multi-target setting.

In this paper, we consider multi-target regression as an example of a multi-target problem and review the current practice of evaluating multi-target regression algorithms. Our goal is to identify key challenges and open problems in evaluating such models, propose possible solutions and start the discussion around the topic.

The rest of the paper is organized as follows. In Section 2, we review recent publications on multi-target regression and provide an overview of the most common approaches towards model evaluation. Their shortcomings are then discussed in Section 3, where we also suggest possible improvements. Additionally, Section 4 inspects widely used benchmark multi-target regression datasets. Finally, Section 5 concludes the paper with a summary of key findings and future research directions.

2 Common practices

A typical machine learning paper introducing a new machine learning algorithm normally contains an evaluation section, where a number of algorithms are run on a set of suitable datasets. The performance of each algorithm on each of the datasets is evaluated using some metric. The obtained scores are compared, sometimes using statistical analysis, in order to make conclusions about the predictive performance of the newly proposed approach. This last step, namely comparison of the models based on experimental results, is not trivial in the presence of multiple target attributes.

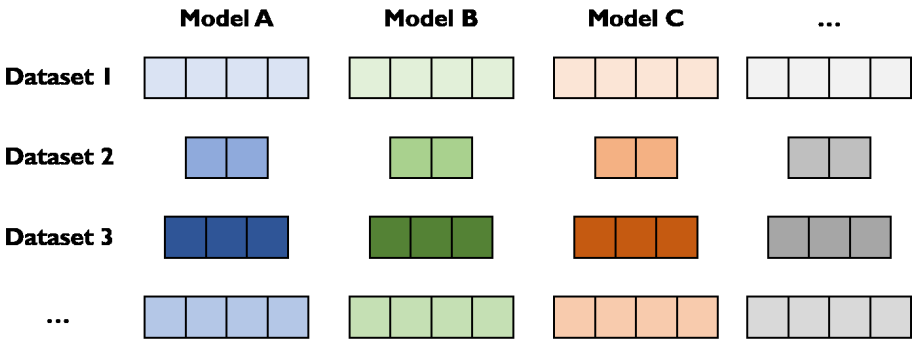


Fig. 1: In the multi-target setting, one obtains several performance scores per dataset (one per each target). It is not trivial to come up with a suitable statistical test to compare such multivariate data. Typical approach is to average scores within a dataset.

To review the current practice, we have analyzed a number of papers that introduce a multi-target regression method and evaluate it by comparing it to a number of competitor techniques. We focused on the works published in the past ten years, and selected representative papers from a variety of authors following different approaches towards tackling a multi-target problem, as well as using different machine learning algorithms as a basis for their methods. The summary of the reviewed papers can be found in Table 1.

Most authors chose for Relative Root Mean Squared Error (RRMSE) to measure the accuracy of the target-specific performance. It is a relative measure that is computed as a ratio of the model's Root Mean Squared Error (RMSE) to that of predicting the average value of the target attribute. The lower the RRMSE, the better.

Other common choices are plain RMSE or Mean Squared Error (MSE). In that case, the authors first standardize the targets so that their values are in the same scales and the corresponding errors are comparable. Additionally, correlation coefficient (CC) between the true and predicted values is sometimes computed.

All estimates are typically obtained via cross-validation (an exception is [19], where a holdout set was used to assess the performance).

Whatever metric is chosen, it leaves one with several scores per dataset, namely one per each target attribute, as illustrated in Figure 1. The common approach is then to average these scores across all targets within each dataset. As a result, a single aggregated performance score per dataset is obtained (e.g., aRRMSE), which makes experimental results look like those in the conventional, single-target case.

The next step is to compare the models among each other. While some papers do not include statistical analysis of the obtained results ([18], [19], [21]), most authors follow the recommendations given by Demšar [4] and perform Friedman

test to check if there are any statistically significant differences between the compared algorithms (or different parameter settings for the same algorithm). If the answer is positive, additional post-hoc tests are performed to find out what these differences are. In addition, average ranks diagrams, which show all the compared algorithms in the order of their average ranks and indicate statistically significant differences, are often plotted to make the results of the statistical analysis easier to comprehend.

Interestingly, however, there seem to be a common uncertainty about which performance scores to run these statistical tests on. Aho et al. [1] state two options, namely:

1. compare the aggregated scores (e.g., aRRMSE), one per dataset;
2. compare individual, target-specific scores (e.g., RRMSE), one per each target in all the datasets.

The authors explicitly state the drawbacks of both approaches. Aggregation across different targets within a dataset (1) is "summing apples and oranges", while comparing target-specific scores (2) is wrong since targets coming from the same dataset are obviously dependent. Nonetheless,"in the absence of a better solution", the authors present results of the statistical analysis for the both options. Similar remarks can be found in [15]. Also in [3] results are reported for the two scenarios[3], while other works ([11], [8], [17], [13]) base their statistical evaluation on the within-dataset averages (1) only.

In the next section, we discuss options (1) and (2) in more detail, as well as suggest possible alternative ways of comparing multi-traget models.

3 Can we do better?

Both approaches to statistical comparison of multi-target models commonly used in practice have a number of drawbacks. We discuss them below.

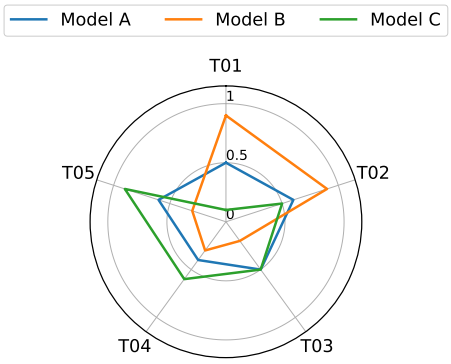
3.1 Why comparing aggregated scores is bad

Apart from not always having a meaningful interpretation, averages are easily affected by the outliers, e.g., when some target is much easier or much more difficult to predict than the others. Excellent performance on an easy target may compensate for the overall bad performance, and vice versa. In addition, when many such targets are strongly correlated, it may appear that the model does very well (badly) on the whole dataset, while actually it is just one task that it did (didn't) manage to learn. Besides, and most importantly, averaging always hides a lot of information. Consider a fictional example given in Figure 2,

[3] Per-target analysis (2) always finds more significant differences in performance of the compared techniques than the per-dataset comparison (1) indicates. This is expected, because statistical test is biased and overly confident in the presence of dependent observations.

(a)

	A	B	C
T1	0.50 (2)	0.90 (3)	0.10 (1)
T2	0.60 (2)	0.90 (3)	0.50 (1)
T3	0.50 (2.5)	0.20 (1)	0.50 (2.5)
T4	0.40 (2)	0.30 (1)	0.60 (3)
T5	0.60 (2)	0.30 (1)	0.90 (3)
	0.52 (2.1)	0.52 (1.8)	0.52 (2.1)



(b)

Fig. 2: (a) In a fictional example where three multi-target models are compared on a dataset with five targets, aRRMSE is the same. However, the target-specific performances are quiet different. Per-target ranks (in brackets) can help highlight the differences. (b) Visualization is key in understanding such differences. Radar plots can be helpful.

where three multi-target models are compared on a dataset with five targets. While the average scores across all targets are the same, models A, B and C perform quite differently. It is not true that all three methods are equally good: depending on the application, one of them can be preferred.

Since within-dataset average scores do not fully reflect the performance of multi-target models, comparisons in terms of such aggregates are not informative, and can even be misleading.

3.2 Why comparing target-specific scores is bad

As has been mentioned in the previous section, the only alternative strategy sometimes used in practice to avoid averaging is to compare target-specific scores across all datasets. As has already been noticed by some researchers, the scores coming from the same dataset are dependent, which violates the assumptions of the Friedman test, commonly applied to compare these scores across the algorithms. Thus, such an approach is not statistically sound and should not be used in practice because the results of the test are not reliable.

Furthermore, even if a statistical test existed that would take the dependencies between performance scores coming from the same dataset into account, it would allow one to compare multi-target models in terms of their performance on a single randomly selected task. Arguably, this is not what we want: one is rather interested in comparing the models based on their *joint* performance on a *set* of related targets.

But is there a statistically sound way to compare a set of target-specific scores as a whole, i.e., without aggregating them into a single score per dataset, yet keeping track of which dataset the scores are coming from?

The task of comparing multivariate samples often arises in practice and has been extensively studied. Examples include nested ANOVA, global statistical tests [14], and "the method of m rankings" [2]. However, to the best of our knowledge, none of the existing approaches are applicable in the multi-target setting, where multiple observations per dataset are dependent, and their number varies across the datasets.

3.3 How can we make the comparisons better

Since there is no suitable statistical test one could run on a table of experimental results such as in Figure 1, some kind of within-dataset aggregation is inevitable. We propose averaging the ranks rather than raw prediction scores. Below, we suggest two ways of introducing ranks.

Target-specific ranks In order to obtain a single aggregated performance score per dataset, one can replace raw target-specific scores by their ranks when averaging across the targets. This will help overcome the issue of non-commensurability of the scores corresponding to different targets, as well as to limit the influence of the outliers.

In the fictional example in Figure 2a, such ranks for a single dataset are given in brackets. While aRRMSEs are the same, the rank of model B is lower than that of models A and C, indicating that this model "wins" more often.

When such ranks are obtained for multiple datasets, statistical tests can be run to compare the models in terms of their average ranks rather then average performance score.

The disadvantage of this approach is that ranks do not capture the magnitude of the differences in performances. Aligned ranks can be introduced that take it into account, but this will make the hypothesis of the statistical test run on these ranks less interpretable.

Pareto ranks Pareto-dominance is an alternative way to compare performances of several algorithms. A model is Pareto-optimal on some dataset if for each target, it yields better prediction accuracy than any other model.

To make Pareto-style comparisons across multiple datasets, Pareto ranks can be introduced as follows. For each dataset, the models that are Pareto-optimal get rank 1. The models that are optimal without considering the models with rank 1, get rank 2, and so on. Once the ranks are computed, statistical tests can be run to check if there is any difference in the Pareto ranks obtained by different models.

The disadvantage of such an approach is that it is quite conservative: improvements across all targets are needed for a model to get a higher rank. It can happen that on some datasets, no model is Pareto-optimal (this is exactly

Dataset	# samples	# features	# targets	% missing	Source
atp1d	337	411	6	0	[15]
atp7d	296	411	6	0	[15]
oes97	334	263	16	0	[15]
oes10	403	298	16	0	[15]
rf1	9125	64	8	0.6	[15]
rf2	9125	576	8	6.8	[15]
scm1d	9803	280	16	0	[15]
scm20d	8966	61	16	0	[15]
edm	154	16	2	0	[10]
sf1	323	10	3	0	[5]
sf2	1066	10	3	0	[5]
jura	359	15	3	0	[7], [15]
wq	1060	16	14	0	[6]
enb	768	8	2	0	[16], [15]
slump	103	7	3	0	[20], [15]
andro	49	30	6	0	[9], [15]
osales	639	413	12	3.8	[15]
scpf	1137	23	3	35.4	[15]

Table 2: Summary of the 18 benchmark datasets used for evaluating multi-target regression models.

the situation in the Figure 2a). In this case, all models get the same rank. If this happens for multiple datasets, no insights can be gained from such a conservative procedure. At the same time, if some algorithm *is* the best in terms of Pareto rank, one can be sure that it outperforms the competitors on all targets, which is not the case when comparison is based on aRRMSE. Every model which is the best in terms of aRRMSE is Pareto-optimal, but the opposite is not true: major improvement on one target can lead to the lowest aRRMSE even if model's performance on the rest of the targets is worse compared to some other methods.

4 Remarks on the benchmark datasets

Datasets used to evaluate the models are as important as the procedures used to draw conclusions about models' performances on them. In this section, we take a closer look at the datasets commonly used to evaluate multi-target regression algorithms.

Illustrating some properties of multi-target methods on toy datasets, or evaluating them on synthetic datasets is unfortunately not common. Only in [19], the methods are evaluated on a synthetic dataset generated using a simulated two-output time series process. This synthetic dataset, however, is not constructed to highlight the differences in the behavior of the compared techniques, but is rather used as an addition to the available real-world data.

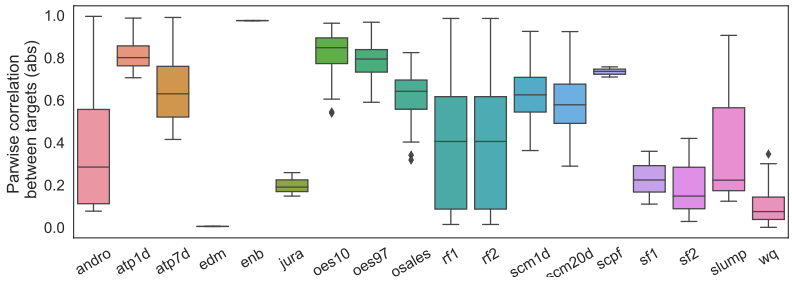


Fig. 3: The distribution of pairwise correlations between targets of benchmark datasets (absolute values are considered to reflect the magnitude of dependencies). Sometimes, all targets are strongly correlated, which can introduce bias in the evaluation process.

In 2016, Spyromitros-Xioufis et al. collected a set of 18 real-world datasets [15] that have been commonly used for evaluation of multi-target regression models since then. These datasets are summarized in Table 2. There are datasets of different sizes in terms of the number of examples, features and targets, which is important to guarantee general evaluation. The datasets are also coming from a variety of different domains such as geology (`jura`), hydrology (`rf`, `wq`, `andro`), astronomy (`sf`) and engineering (`edm`, `enb`, `slump`), as well as e-commerce (`scpf`), sales (`atp`, `osales`), economics (`oes`) and management (`scm`). However, while datasets seem diverse at first glance, there are two aspects that need to be taken into account when using them to evaluate multi-target models.

First, one can notice that some of them come in pairs, e.g., `sf1` and `sf2`, etc. This is because, in some cases, separate datasets were created for the same type of data collected, for instance, in different years. Such datasets are thus very similar, and the algorithms are likely to demonstrate similar performance on them, which can add bias to the evaluation process. Besides, performance scores, both target-specific and aggregated, coming from such similar datasets are also dependent, which is a problem for statistical tests.

The second point worth discussing is the magnitude of dependency between the targets in a single dataset. In Figure 3, we plot the distribution of the absolute values of the pairwise correlations between the targets per dataset. One can notice that for some datasets such as `atp1d`, `oes10` and `oes97`, these values are rather high. An extreme case is the `enb` dataset, where the only two targets are perfectly correlated. An opposite situation can be observed in `enb`, where there is no correlation between the targets. Of course, this is not 'wrong' per se since the data is coming from the real-world applications and reflects the phenomena that occur in practice. However, these aspects should be taken into account during the evaluation process, since predicting linearly dependent targets is easier than those with more complex inter-dependency. One should therefore not make overly

confident statements about the performance of a multi-target model when testing on the datasets with very correlated targets.

5 Conclusions

Multi-target regression, which is a special case of multi-target prediction where several numerical targets are predicted simultaneously, is an actively developing field with diverse real-world applications. A lot of methods to tackle multi-target regression tasks are being proposed by the researchers in the field. In this paper, we have addressed the problem of evaluating multi-target regression models, which is crucial to better understand and modify the existing techniques, as well as to successfully develop new ones. Our analysis of the recent papers publishing the results of several multi-target regression methods over multiple datasets has shown that many authors are unsure about the correct way of making such comparisons.

We argue that comparing multi-target models in terms of averaged scores leaves no possibility to fully understand and meaningfully discuss strengths and weaknesses of different approaches. Besides, it does not help the practitioners to choose an appropriate model to solve a real-world multi-target regression problem. However, we conclude that to run statistical test on experimental results, aggregation is inevitable since no existing test is suitable for comparing multi-target models. We therefore propose two ways on how ranks can be used instead of raw performance scores to overcome such shortcomings of simple averaging as non-commensurability of the target-specific scores and sensitivity to outliers. Note that such ranks, just like any other aggregation, hide a lot of information about models' behavior, and should be reported along with other metrics (e.g., min and max of target-specific errors, median error, etc.).

Simply plotting the per-target results can help highlight the differences between the models. Radar plot, shown for a fictional example in Figure 2b, is a good example of such visualization. Besides, it is worth to explore why some models perform well on one set of targets and others on another one, as it gives a deeper understanding of the behavior of the model. Such analysis is almost not happening in practice. One example can be found in [17].

In addition, we also inspect benchmark multi-target regression datasets and claim that more diverse datasets are needed to improve the evaluation process. In the absence of more real-world data, one solution is to use artificially generated datasets[4]. Wisely created, such datasets are useful not only to compare the overall predictive performance of multi-target approaches, but also to explore and understand the behavior of individual algorithms in-depth.

References

1. Aho, T., Ženko, B., Džeroski, S., Elomaa, T.: Multi-target regression with rule ensembles. Journal of Machine Learning Research **13**(Aug), 2367–2407 (2012)

[4] An attempt to create such toy benchmarks can be seen in [12].

2. Benard, A.p., vanElteren, P.: A generalization of the method of m rankings. Indagationes Mathematicae **1**(5), 358–369 (1953)
3. Breskvar, M., Kocev, D., Džeroski, S.: Ensembles for multi-target regression with random output selections. Machine Learning **107**(11), 1673–1709 (2018)
4. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)
5. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
6. Džeroski, S., Demšar, D., Grbović, J.: Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence **13**(1), 7–17 (2000)
7. Goovaerts, P.: Geostatistics for natural resources evaluation. Oxford University Press on Demand (1997)
8. Hadavandi, E., Shahrabi, J., Shamshirband, S.: A novel boosted-neural network ensemble for modeling multi-target regression problems. Engineering Applications of Artificial Intelligence **45**, 204–219 (2015)
9. Hatzikos, E.V., Tsoumakas, G., Tzanis, G., Bassiliades, N., Vlahavas, I.: An empirical study on sea water quality prediction. Knowledge-Based Systems **21**(6), 471–478 (2008)
10. Karalič, A., Bratko, I.: First order regression. Machine learning **26**(2-3), 147–176 (1997)
11. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recognition **46**(3), 817–833 (2013)
12. Mastelini, S.M., Santana, E.J., da Costa, V.G.T., Barbon, S.: Benchmarking multi-target regression methods. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). pp. 396–401. IEEE (2018)
13. Melki, G., Cano, A., Kecman, V., Ventura, S.: Multi-target support vector regression via correlation regressor chains. Information Sciences **415**, 53–69 (2017)
14. O'Brien, P.C.: Procedures for comparing samples with multiple endpoints. Biometrics pp. 1079–1087 (1984)
15. Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., Vlahavas, I.: Multi-target regression via input space expansion: treating targets as inputs. Machine Learning **104**(1), 55–98 (2016). https://doi.org/10.1007/s10994-016-5546-z, http://dx.doi.org/10.1007/s10994-016-5546-z
16. Tsanas, A., Xifara, A.: Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings **49**, 560–567 (2012)
17. Tsoumakas, G., Spyromitros-Xioufis, E., Vrekou, A., Vlahavas, I.: Multi-target regression via random linear target combinations. In: Joint european conference on machine learning and knowledge discovery in databases. pp. 225–240. Springer (2014)
18. Tuia, D., Verrelst, J., Alonso, L., Pérez-Cruz, F., Camps-Valls, G.: Multioutput support vector regression for remote sensing biophysical parameter estimation. IEEE Geoscience and Remote Sensing Letters **8**(4), 804–808 (2011)
19. Xu, S., An, X., Qiao, X., Zhu, L., Li, L.: Multi-output least-squares support vector regression machines. Pattern Recognition Letters **34**(9), 1078–1084 (2013)
20. Yeh, I.C.: Modeling slump flow of concrete using second-order regressions and artificial neural networks. Cement and Concrete Composites **29**(6), 474–480 (2007)
21. Zhen, X., Yu, M., He, X., Li, S.: Multi-target regression via robust low-rank learning. IEEE transactions on pattern analysis and machine intelligence **40**(2), 497–504 (2017)