ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Problem Transformation Methods with Distance-Based Learning for Multi-Target Regression

Joonas Hämäläinen, Tommi Kärkkäinen

University of Jyvaskyla, Faculty of Information Technology,
P.O. Box 35, FI-40014 University of Jyvaskyla, Finland

Abstract. Multi-target regression is a special subset of supervised machine learning problems. Problem transformation methods are used in the field to improve the performance of basic methods. The purpose of this article is to test the use of recently popularized distance-based methods, the minimal learning machine (MLM) and the extreme minimal learning machine (EMLM), in problem transformation. The main advantage of the full data variants of these methods is the lack of any meta-parameter. The experimental results for the MLM and EMLM show promising potential, emphasizing the utility of the problem transformation especially with the EMLM.

1 Introduction

Multi-Target Regression [1, 2] (MTR) refers to a machine learning problem where multiple real-valued target variables are predicted at once with a given input. Prime example of MTR is the Multi-Target Classification (MTC) problem where an input vector is linked to the membership in multiple different classes. MTR and MTC are subfields of Multi-Output Learning which has been an active field of research recently [3]. Although MTR has an increasing number of interesting applications, the most focus in the field has been in MTC problems [2].

If a regression method, e.g., the support vector machine, does not pose an inherent capability to build a model for multi-targets, the so-called Problem Transformation [1, 2] (PT) methods can be applied for the extension. The most common and straightforward problem transformation method, the Single Target (ST) method, builds a separate model for each target variable. Lately, the popular problem transformation methods from MTC were adapted to MTR [1]. Based on the extensive experimental and theoretical evaluation in [1], the Stacked Single-Target (SST) and the Ensemble Regressor Chains (ERC) combined with the cross-validation based extension of the input dataset were concluded to improve the prediction accuracy compared to the ST approach. The common idea here is to form first stage models with the ST approach and then use the predictions of these models to form an extended input dataset (meta-dataset). This enlarged data is then mapped back to the original targets with a second stage ST model. Naturally, the base regressor that is used in the ST approach in SST and ERC affects to the computational cost and accuracy of the whole extended model.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Lately, two distance-based supervised learning methods have been proposed for classification and regression: the Minimal Learning Machine (MLM) [4] and the Extreme Minimal Learning Machine (EMLM) [5]. Both of these models possess attractive characteristics that suggest testing their use as the base regressor for SST and ERC. Similar to the Extreme Learning Machine, MLM and EMLM have simple formulation, one hyperparameter, and good trade-off between the accuracy and computational cost. Differently to ELM, tuning hyperparameter for MLM and EMLM is mostly an issue to balance the computational cost and the generalization accuracy [6, 5]. Therefore, nonparametric full data variants of the MLM and EMLM methods enable their straightforward use in SST and ERC, especially because one of the main reasons for the popularity of these formulations in MTC is their simplicity [1].

2 Methods

Here we briefly describe the basics of problem transformation and distance-based regression methods.

2.1 Problem Transformation

In MTR, the goal is to learn a mapping $\mathbf{f}$ between the input data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and the output data $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^M$, $\mathbf{y}_i \in \mathbb{R}^L$ and $L > 1$. The ST approach builds L models for the individual outputs $\mathbf{Y}_j = \{(\mathbf{y}_i)_j\}_{i=1}^N$. Let $(\tilde{\mathbf{y}}_i)_j \in \mathbb{R}$ be the predicted value for the jth ST model $\mathbf{f}_{ST_j}$ corresponding to the input vector $\mathbf{x}_i$. The meta-dataset $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^N$ is then built for SST [1] so that $\tilde{\mathbf{x}}_i = [(\mathbf{x}_i)_1, \ldots, (\mathbf{x}_i)_M, (\tilde{\mathbf{y}}_i)_1, \ldots, (\tilde{\mathbf{y}}_i)_L]$. In the second learning stage, ST is applied to learn a mapping $\tilde{\mathbf{f}}_{ST_j}$ between $\tilde{\mathbf{X}}$ and $\mathbf{Y}_j$ for each j. In the prediction phase of the SST, for a new input vector $\hat{\mathbf{x}}$, the first stage prediction $\hat{\mathbf{y}}$ is computed with the models $\{\mathbf{f}_{ST_j}\}_{j=1}^L$ to create the extended input vector $\hat{\mathbf{x}}' = [\hat{\mathbf{x}}, \hat{\mathbf{y}}]$. This is then given to the second stage models $\{\tilde{\mathbf{f}}_{ST_j}\}_{j=1}^L$ to get the actual prediction.

Similarly to SST, the ERC [1] method also forms an extended input dataset from the predictions, but with an iterative fashion. Let us assume that variables of target set $\mathbf{Y}$ have been randomly ordered. In the first learning stage of the ERC, the ST method is used to form a mapping $\mathbf{f}_{ST_1}$ for $\mathbf{X}$ and $\mathbf{Y}_1$ with the corresponding predictions $\{(\tilde{\mathbf{y}}_i)_1^{t=0}\}_{i=1}^N$. Then meta-dataset for each of next iterations $t = 1, ..., L-1$ is $\tilde{\mathbf{X}}^t = \{\tilde{\mathbf{x}}_i^t\}_{i=1}^N$, where $\tilde{\mathbf{x}}_i^t = [(\mathbf{x}_i)_1, \ldots, (\mathbf{x}_i)_M, (\tilde{\mathbf{y}}_i)_1^0, \ldots, (\tilde{\mathbf{y}}_i)_{t-2}^{t-1}]$. In each of the iterations $t = 1, ..., L-1$, the mapping $\mathbf{f}_{ST_{t+1}}$ between $\tilde{\mathbf{X}}^t$ and $\mathbf{Y}_{t+1}$ is built which gives the corresponding predictions $\{(\tilde{\mathbf{y}}_i)_{t+1}\}_{i=1}^N$. Because the performance of the ERC is sensitive to the order of the chains [3, 1], this training procedure is repeated q times with different chain orders. Training phase produces q different chains, where each is constructed of L different sized models. Prediction for new input vector is computed by following the same chain order that was used in the training. Each chain gives chain's prediction from the output of the last model $\mathbf{f}_{ST_L}$ and the actual prediction is given by the average

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

of these predictions.

According to [1], using predictions from the CV sets is recommended for building the meta-datasets, because the CV-based values are more similar to values in the prediction phase than the training targets. Note that the CV models are only used in the training phase to build the meta-dataset. An explanation for improved generalization accuracy of the SST and ERC approaches compared to the ST approach could be reasoned by the modeling capability of the dependencies between the target variables. Another interpretation is that SST and ERC methods out-weight variance to decrease bias [1].

2.2 Regression via Distance-Based Learning: MLM and EMLM

In the training phase of MLM [4], subsets of K reference points for the input data $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^{K} \subseteq \mathbf{X}$ and the output data $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^{K} \subseteq \mathbf{Y}$ are selected. Then, distance matrices $\mathbf{D_x}$ and $\mathbf{D_y}$ are computed such that $\mathbf{D}_{\mathbf{x}(i,k)} = d(\mathbf{x}_i, \mathbf{r}_k)$ and $\mathbf{D}_{\mathbf{y}(i,k)} = d(\mathbf{y}_i, \mathbf{t}_k)$ where $i = 1, ..., N$, $j = 1, ..., K$, and $d(\cdot, \cdot)$ denotes the Euclidean distance. Finally, the regression model with the distance matrices is determined by the solution of the Ordinary Least-Squares (OLS) solution $\mathbf{B} = (\mathbf{D_x}^T\mathbf{D_x})^{-1}\mathbf{D_x}^T\mathbf{D_y}$ when $K < N$. For the Full MLM case, for $K = N$, the solution is simply $\mathbf{D_x}^{-1}\mathbf{D_y}$ when the inverse exists [6].

In the prediction phase, distances to the input reference points of the new input $\hat{\mathbf{x}}$ are computed and the output space distances $\boldsymbol{\delta} \in \mathbb{R}^K$ are computed from the distance matrix regression model as $\boldsymbol{\delta} = [d(\hat{\mathbf{x}}, \mathbf{r}_1), ..., d(\hat{\mathbf{x}}, \mathbf{r}_K)]\mathbf{B}$. The actual prediction $\hat{\mathbf{y}}$ is then obtained by solving a squared stress optimization problem [4, 5]. In the case of a single target regression, the solution of this problem is given by the roots of the cubic equation [7]. Hence, this MLM variant is also referred as Cubic MLM (C-MLM).

The EMLM method is a simplification of MLM as proposed and tested in [8, 5]. It builds a distance-based regression model directly using the input-distance matrix $\mathbf{D_x}$ as the kernel. Unique solvability (positive definiteness) of the resulting OLS problem can be guaranteed by using Tikhonov regularization.

3 Experiments

In this section, we report the experimental results with five MTR datasets from `http://mulan.sourceforge.net/datasets-mtr.html`. Similar to [1], the performance of the methods was evaluated with the average Relative Root Mean Squared Error (aRRMSE) and the relative improvement of the aRRMSE.

3.1 Experimental setup

Characteristics of the selected datasets are shown in Table 1 (see [1] for detailed description of the datasets). All features were first minmax-scaled into $[0, 1]$. Rows with missing value were removed from the RF1 dataset.We randomly partitioned the datasets into training (80%) and testing sets (20%). We repeated

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Name	N	M	L
JURA	359	15	3
WQ	1060	16	14
RF1	9125	64	8
SCM20D	8966	61	16
SCM1D	9803	280	16

Table 1: Datasets

Method name	Core PT method	Base regressor	Metadata based on
ST-MLM	ST	Full C-MLM	-
SST-MLM	SST	Full C-MLM	DOB-SCV
ERC-MLM	ERC	Full C-MLM	DOB-SCV
ST-EMLM	ST	Full EMLM	-
SST-EMLM	SST	Full EMLM	DOB-SCV
ERC-EMLM	ERC	Full EMLM	DOB-SCV

Table 2: Summary of the MLM and EMLM variants for MTR.

these divisions 30 times for JURA, WQ, and RF1, and 10 times for SCM20D and SCM1D. Results for the aRRMSE error were averaged over the repetitions.

All the methods were implemented and tested in MATLAB R2018b environment. All the experiments were run on an eight-node computing cluster, where each node was equipped with eight-core Intel Xeon CPU E7-8837 with 128 GB memory. Description of the methods is summarized in Table 2. In the SST and ERC variants, instead of the standard CV, we used the Distribution Optimally Balanced Stratified Cross-Validation (DOB-SCV) [9] method with 10 folds to construct the meta-dataset. For ERC-MLM and ERC-EMLM, we generated 10 different random chains for the datasets where $L! \geq 10$ and $L!$ different random chains otherwise. The ERC-MLM method was only applied to the JURA and WQ datasets due to its high computational cost.

RRMSE for a target variable j is given by

$$\mathrm{RRMSE}_j = \sqrt{\frac{\sum_{i=1}^{N_{test}}((\tilde{\mathbf{y}}_i)_j - (\mathbf{y}_i)_j)^2}{\sum_{i=1}^{N_{test}}(\bar{\mathbf{Y}}_j) - (\mathbf{y}_i)_j)^2}}, \tag{1}$$

where N_{test} is the size of the test set, $\tilde{\mathbf{y}}$ is the predicted target, $\mathbf{y}$ is the ground truth target, and $\bar{\mathbf{Y}}_j$ is the mean target value in the training dataset for the variable j. Then, the aRRMSE is given by $\mathrm{aRRMSE} = \frac{1}{L}\sum_{j=1}^{L}\mathrm{RRMSE}_j$. Finally, the relative improvement of aRRMSE is defined as [1] $\mathrm{RI} = \frac{aRRMSE_{ST}}{\mathrm{aRRMSE}_m}$, where aRRMSE_{ST} is the aRRMSE for an ST method and aRRMSE_m is the aRRMSE for a method m. RI values can be used to estimate how much the overall accuracy is improved with respect to the ST method. RI-values larger 1 indicate that the accuracy is improved and the values smaller than 1 indicate that the accuracy is degenerated.

3.2 Results

Results for aRRMSE are summarized in Table 3. The SST and ERC can improve the overall accuracy of MLM and EMLM. Note that all the MLM and EMLM variants using the same core PT method seem to have similar aRRMSE. The SST variants give the smallest average aRRMSE for all other datasets except for WQ, where the ERC variants are clearly better.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Dataset	ST-MLM	SST-MLM	ERC-MLM	ST-EMLM	SST-EMLM	ERC-EMLM
JURA	0.37698	0.36820	0.37295	0.37603	**0.36508**	0.37238
QW	0.90623	0.90732	**0.85139**	0.90912	0.90859	0.85376
RF1	0.20381	0.10337	-	0.20205	**0.10296**	0.20072
SCM20D	0.09349	**0.08649**	-	0.09741	0.08843	0.08950
SCM1D	0.07243	**0.07028**	-	0.07345	0.07069	0.07159

Table 3: Average aRRMSE in the testing set.

Next we focus on analyzing the RI measure which is more informative than the average aRRMSE. Results for RI are shown in Figure 1. For the RF1 dataset, we computed RI separately without the second target variable, because aRRMSE was highly dominated by the second target variable's RRMSE. The second target variable of RF1 is highly discrete compared to the other target variables. RI results for the second target variable are shown separately in Figure 1.

Based on Figure 1, SST seems to improve the relative accuracy of MLM and EMLM more than the ERC approach. Significantly, all the SST and ERC variants improve aRRMSE in each repetition for the two largest datasets SCM20D and SCM1D. In spite of the fact that SST improves average aRRMSE more than ERC for the JURA and RF1 datasets, the ERC variants have smaller degeneration of the accuracy in rare cases when it happens. Based on the results for SCM20D and SCM1D in Figure 1, the accuracy of EMLM with SST is more improved than the accuracy of MLM with SST. The SST-EMLM variant seems very promising for further studies, because SST have much smaller computational cost than ERC [1] and EMLM have also much smaller cost than MLM [5]. Therefore, the SST-EMLM is the fastest method from the SST and ERC variants.

4 Conclusions

Previous works in multi-target classification have shown the promising potential of the problem transformation methods. However, in multi-target regression the subject of problem transformation is rarely studied. In this paper, we evaluated the SST and ERC problem transformation methods with the lately proposed distance-based regression methods, MLM and EMLM, in the multi-target regression problems. The results demonstrate the potential of the problem transformation with the distance-based regression. The SST and ERC variants of the MLM and EMLM methods can improve the overall generalization accuracy compared to the single-target MLM and EMLM methods. In the future work, addressing issues of the problem transformation method selection and the target variable selection based on the characteristics of the data could be studied. Because of the efficiency and simplicity of the nonparametric SST-EMLM method, it is also an interesting subject for further studies.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

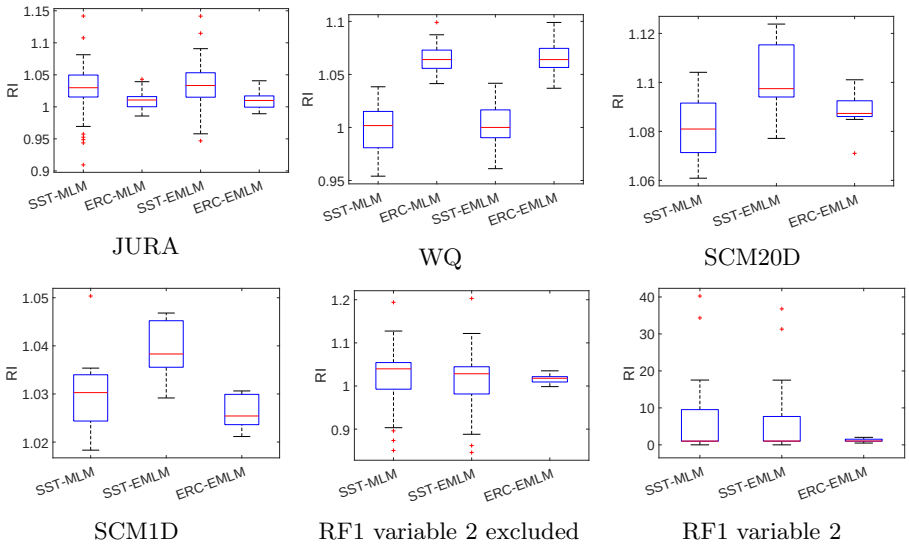

Fig. 1: Results for the relative improvement RI.

References

[1] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98, 2016.

[2] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

[3] Donna Xu, Yaxin Shi, Ivor W Tsang, Yew-Soon Ong, Chen Gong, and Xiaobo Shen. A survey on multi-output learning. *arXiv preprint arXiv:1901.00248*, 2019.

[4] Amauri Holanda de Souza Junior, Francesco Corona, Guilherme A. Barreto, Yoan Miche, and Amaury Lendasse. Minimal learning machine: A novel supervised distance-based approach for regression and classification. *Neurocomputing*, 164:34–44, 2015.

[5] Tommi Kärkkäinen. Extreme minimal learning machine: Ridge regression with distance-based basis. *Neurocomputing*, 342:33–48, 2019.

[6] Joonas Hämäläinen, Alisson SC Alencar, Tommi Kärkkäinen, César LC Mattos, Amauri H Souza Júnior, and João PP Gomes. Minimal learning machine: Theoretical results and clustering-based reference point selection. *arXiv preprint arXiv:1909.09978*, 2019.

[7] Diego P. P. Mesquita, João P. P. Gomes, and Amauri H. Souza Junior. Ensemble of efficient minimal learning machines for classification and regression. *Neural Processing Letters*, pages 1–16, 2017.

[8] Tommi Kärkkäinen. Extreme minimal learning machine. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 237–242. ESANN, 2018.

[9] J. G. Moreno-Torres, J. A. Sáez, and F. Herrera. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312, 2012.