DEEP METRIC LEARNING-BASED SEMI-SUPERVISED REGRESSION WITH ALTERNATE LEARNING

Adina Zell, Gencer Sumbul, Begüm Demir

Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany

ABSTRACT

This paper introduces a novel deep metric learning-based semi-supervised regression (DML-S2R) method for parameter estimation problems. The proposed DML-S2R method aims to mitigate the problems of insufficient amount of labeled samples without collecting any additional samples with target values. To this end, the proposed DML-S2R method is made up of two main steps: i) pairwise similarity modeling with scarce labeled data; and ii) triplet-based metric learning with abundant unlabeled data. The first step aims to model pairwise sample similarities by using a small number of labeled samples. This is achieved by estimating the target value differences of labeled samples with a Siamese neural network (SNN). The second step aims to learn a triplet-based metric space (in which similar samples are close to each other and dissimilar samples are far apart from each other) when the number of labeled samples is insufficient. This is achieved by employing the SNN of the first step for triplet-based deep metric learning that exploits not only labeled samples but also unlabeled samples. For the end-to-end training of DML-S2R, we investigate an alternate learning strategy for the two steps. Due to this strategy, the encoded information in each step becomes a guidance for learning the other step. The experimental results confirm the success of DML-S2R compared to the state-of-the-art semi-supervised regression methods. The code of the proposed method is publicly available at `https://git.tu-berlin.de/rsim/DML-S2R`.

Index Terms— Semi-supervised regression, parameter estimation, metric learning, deep learning.

1. INTRODUCTION

The accurate estimation of parameters from specific data (e.g., estimation of carbon monoxide concentration in the environment, estimation of forest parameters, estimation of glucose concentration in diabetic patients, etc.) is an important research field in machine learning and pattern recognition [1]. The use of regression methods, which aim at learning functional relations between a set of variables and corresponding target values, is an effective way for parameter estimation problems. Accordingly, several regression methods (e.g., random forest regression [2], support vector regression [1], deep learning based regression [3, 4], etc.) have been introduced in the literature. The success of these methods depends on the quantity and quality of training samples for which the respective target values are available (i.e., labeled samples). The quantity of training samples depends on the number of available labeled samples, while the quality of training samples depends on their capability to represent the real sample distribution. A small amount of labeled samples with insufficient quality can lead to inadequate modeling of the regression task. However, collecting labeled samples is often expensive and complex as producing reference measures may require significant time [1]. To address this issue, semi-supervised regression (SSR) methods, which aim at jointly using the information of both labeled and unlabeled samples in the learning phase of the regression algorithm, have been introduced [5–7].

In recent years, deep metric learning (DML) has been utilized in the framework of SSR. DML operates on sample tuples (e.g., triplets) to learn a metric space, in which similar samples are mapped close to each other and dissimilar samples are mapped apart from each other. In [8], a metric-learning based SSR method that utilizes a Siamese neural network (SNN) with contrastive loss function is introduced to learn a pairwise metric space. In this method, dissimilar and similar pairs, which are constructed from both labeled and unlabeled samples, are utilized for DML. To define sample similarity, the absolute differences of target values are exploited with a thresholding strategy for labeled samples. The similarity of unlabeled samples to labeled samples is estimated based on Euclidean distances between data points, for which thresholding is applied. Once the SNN is trained with dissimilar and similar sample pairs, the target value estimation of samples is performed on the learnt metric space by k-nearest neighbors (k-NN) algorithm. This method only considers the pairwise similarities, which may not be sufficient in accurately learning a metric space. In [9], a SSR method that exploits a long short-term memory (LSTM) based SNN is proposed for continuous emotion recognition. This method applies a two-stage training. In the first stage, the SNN is trained with log-ratio loss function (which operates on sample triplets) by using only labeled samples. Then, the SNN is used to generate pseudo-labels for unlabeled samples. In the second stage, the SNN is trained with mean squared error loss function by using pseudo-labels. This method utilizes only labeled samples for learning metric space in the first stage. This may lead to inaccurate characterization of sample similarities in the case of availability of small-sized labeled training sets. To address the limitations of above-mentioned methods, we introduce a novel **D**eep **M**etric **L**earning-based **S**emi-**S**upervised **R**egression (DML-S2R) method for parameter estimation.

2. PROPOSED DML-S2R METHOD

Let $\mathcal{S} = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^N$ be a set of N labeled samples, where $\boldsymbol{x}_i^l$ is the ith labeled sample and $\boldsymbol{y}_i^l$ is its corresponding target value. Let $\mathcal{U} = \{\boldsymbol{x}_j^u\}_{j=1}^M$ be the set of M unlabeled samples, where $\boldsymbol{x}_j^u$ is the jth sample, for which the corresponding target value is unknown. The training set $\mathcal{T}$ consists of the labeled sample set $\mathcal{S}$ and the unlabeled sample set $\mathcal{U}$ ($\mathcal{T} = \mathcal{S} \cup \mathcal{U}$). We assume that N is significantly lower than M ($N << M$).

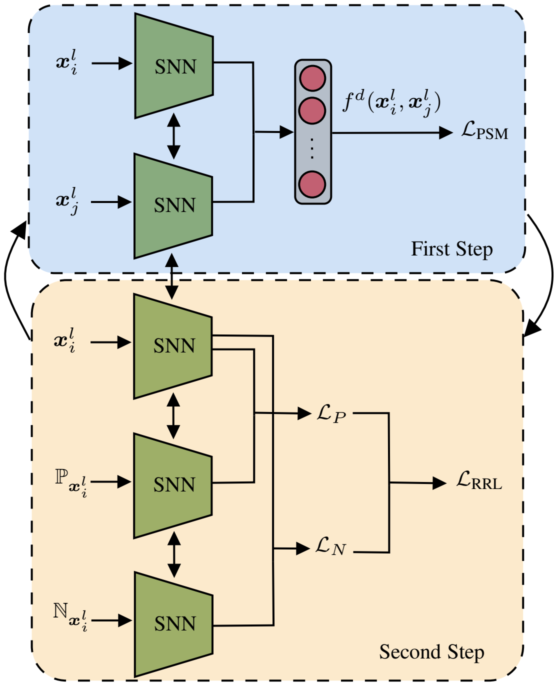
The proposed DML-S2R method aims to learn a metric space, in which similar samples are located close to each other, by effectively exploiting abundant unlabeled data together with scarce labeled data. This is achieved by two main steps: i) pairwise similarity modeling with scarce labeled data; and ii) triplet-based metric learning with abundant unlabeled data. For the end-to-end training of DML-S2R, we investigate an alternate learning strategy for the two steps. Fig. 1 shows a general overview of the proposed method, which is explained in detail in the following subsections.

2.1. Pairwise Similarity Modeling with Scarce Labeled Data

The first step aims to model the pairwise sample similarity by using a small amount of labeled samples. For this purpose, one could exploit contrastive loss function that requires the selection of similar and dissimilar pairs based on the target values of training samples. However, in the framework of regression problems, it is challenging to define the boundaries between the target values of similar/dissimilar samples [10]. To overcome this problem, we redefine the regression problem. Instead of learning a function that estimates the target value of each sample, DML-S2R learns a function f^d which estimates the target value differences of two labeled samples as:

$$f^d(\boldsymbol{x}_i^l, \boldsymbol{x}_j^l) = \boldsymbol{y}_i^l - \boldsymbol{y}_j^l. \tag{1}$$

While learning to estimate target value differences, DML-S2R is also enforced to model the sample similarity for the accurate prediction of the differences. Due to this definition, instead of using N samples in the first step, proposed DML-S2R exploits $N(N-1)$ pairs of samples (i.e., all the possible pairs from labeled sample set). This mitigates the limitations of labeled data scarcity, and thus over-fitting. For f^d, we utilize a SNN, which contains two identical sub-networks with shared weights. Two sub-networks of the SNN takes an input pair and provides the sample features, which are concatenated to form the feature associated to the input pair. Then, this feature is fed into a fully connected layer (FC) that directly estimates the target value differences of sample pairs. Let $\boldsymbol{z}_{ij}$ be the target value difference of $\boldsymbol{x}_i^l$ and $\boldsymbol{x}_j^l$ while $i \neq j$. To learn the model parameters of f^d, we propose the pairwise similarity modelling loss function $\mathcal{L}_{\text{PSM}}$ as follows:



$\boldsymbol{x}_i^l$
$\boldsymbol{x}_j^l$
SNN
SNN
$f^d(\boldsymbol{x}_i^l, \boldsymbol{x}_j^l)$
$\mathcal{L}_{\text{PSM}}$
First Step
$\boldsymbol{x}_i^l$
$\mathbb{P}_{\boldsymbol{x}_i^l}$
$\mathbb{N}_{\boldsymbol{x}_i^l}$
SNN
SNN
SNN
$\mathcal{L}_P$
$\mathcal{L}_N$
$\mathcal{L}_{\text{RRL}}$
Second Step

Fig. 1: Illustration of the proposed DML-S2R method.

$$\mathcal{L}_{\text{PSM}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{1}_{[i \neq j]} (\boldsymbol{z}_{ij} - f^d(\boldsymbol{x}_i^l, \boldsymbol{x}_j^l))^2, \tag{2}$$

where $\mathbb{1}$ is the indicator function. Once the model parameters of the SNN is learnt with (2), its feature space, which encodes the pairwise sample similarity, forms the basis for the second step.

2.2. Triplet-Based Metric Learning with Abundant Unlabeled Data

The second step aims to learn a metric space (where similar samples are located close to each other) when the number of labeled samples is limited. This is achieved by triplet-based DML that takes into account not only labeled samples but also unlabeled samples. Triplet-based DML requires sample triplets for the characterization of a metric space. Accordingly, we convert the SNN of the first step to the triplet-based SNN by replicating one of its identical sub-networks three times without changing weights, and thus make it appropriate for triplet-based DML. A standard triplet consists of a sample anchor and a positive sample, which is similar to the an-

Algorithm 1: Positive-negative set selection of an anchor for the proposed DML-S2R method

```
Input  : $\boldsymbol{x}^l_a, \mathcal{U}, f^d, k$
Output: $\mathbb{P}_{\boldsymbol{x}^l_a}, \mathbb{N}_{\boldsymbol{x}^l_a}$
$\mathbb{P}_{\boldsymbol{x}^l_a} = \emptyset, \mathbb{N}_{\boldsymbol{x}^l_a} = \emptyset$
while $|\mathbb{P}_{\boldsymbol{x}^l_a}| \leq k$ do
    $\underset{\boldsymbol{x}^u_i \in \mathcal{U} \backslash \mathbb{P}_{\boldsymbol{x}^l_a}}{\text{argmin}} \; f^d(\boldsymbol{x}^l_a, \boldsymbol{x}^u_i) \rightarrow \mathbb{P}_{\boldsymbol{x}^l_a}$ (Positive set selection)
end
while $|\mathbb{N}_{\boldsymbol{x}^l_a}| \leq k$ do
    $\underset{\boldsymbol{x}^u_i \in \mathcal{U} \backslash \mathbb{N}_{\boldsymbol{x}^l_a}}{\text{argmax}} \; f^d(\boldsymbol{x}^l_a, \boldsymbol{x}^u_i) \rightarrow \mathbb{N}_{\boldsymbol{x}^l_a}$ (Negative set selection)
end
return $\mathbb{P}_{\boldsymbol{x}^l_a}, \mathbb{N}_{\boldsymbol{x}^l_a}$
```

chor, and a negative sample, which is dissimilar to the anchor. For the construction of sample triplets, anchor samples are selected from the labeled set $\mathcal{S}$ and positive and negative samples are selected from the unlabeled set $\mathcal{U}$. To effectively exploit abundant unlabeled data, we create a set of positive and negative samples per anchor that leads a faster convergence compared having only one positive and one negative sample per anchor [11]. For an anchor $\boldsymbol{x}^l_a \in \mathcal{S}$, the set of k positive samples $\mathbb{P}_{\boldsymbol{x}^l_a}$ and the set of k negative samples $\mathbb{N}_{\boldsymbol{x}^l_a}$ are selected with f^d from the first step based on their estimated target value differences with the anchor. In detail, all the possible pairs between the anchor $\boldsymbol{x}^l_a$ and each unlabeled sample $\boldsymbol{x}^u_i \in \mathcal{U}$ are created. Then, the target value difference of each pair is estimated by using f^d from the first step. The k unlabeled samples having the smallest difference with the anchor are selected for $\mathbb{P}_{\boldsymbol{x}^l_a}$, while the k unlabeled samples having the highest difference with the anchor are selected for $\mathbb{N}_{\boldsymbol{x}^l_a}$. The positive-negative set selection procedure is shown in Algorithm 1. After the selection of the positive and negative sets for all anchor samples, we employ the ranked list loss function [11] as follows:

$$\begin{aligned} w(\boldsymbol{x}^l_a, \boldsymbol{x}^u_j) &= \exp(\tau(d(\boldsymbol{x}^l_a, \boldsymbol{x}^u_j) - (\alpha - m))), \\ \mathcal{L}_P(\boldsymbol{x}^l_a, \mathbb{S}) &= \sum_{\boldsymbol{x}^u_j \in \mathbb{S}} \frac{w(\boldsymbol{x}^l_a, \boldsymbol{x}^u_j)}{\sum_{\boldsymbol{x}^u_j \in \mathbb{S}} w(\boldsymbol{x}^l_a, \boldsymbol{x}^u_j)} \mathcal{L}_m(\boldsymbol{x}^l_a, \boldsymbol{x}^u_j), \\ \mathcal{L}_{\text{RLL}} &= \frac{1}{2N} \sum_{i=1}^{N} \mathcal{L}_P(\boldsymbol{x}^l_a, \mathbb{P}_{\boldsymbol{x}^l_a}) + \mathcal{L}_P(\boldsymbol{x}^l_a, \mathbb{N}_{\boldsymbol{x}^l_a}), \end{aligned} \tag{3}$$

where τ is the temperature parameter, α is the negative sample boundary, m is the margin parameter, d measures the Euclidean distance between two samples in the feature space and $\mathcal{L}_m$ is the margin loss function. $\mathcal{L}_{\text{RLL}}$ pulls a set of positive samples closer than the set of negatives by the margin m on the feature space of the SNN.

It is worth noting that the effectiveness of each step in DML-S2R depends on each other. Inaccurate learning of f^d in the first step leads to incorrect selection of positive-negative sets in the second step. If the metric space is not accurately learned in the second step, the weights of the SNN can not be effectively learnt in the first step due to a small number of labeled samples. This prevents to utilize standard joint learning strategies for DML-S2R. To accurately learn both steps, we investigate an alternate learning strategy, in which the SNN is trained for both steps within the consecutive training epochs. In detail, the whole learning procedure starts with training the SNN one epoch for the first step while minimizing $\mathcal{L}_{\text{PSM}}$. Then, the SNN is trained one epoch with all the anchor samples and the associated positive-negative sets while minimizing $\mathcal{L}_{\text{RRL}}$. Training continues while alternating between the two steps until convergence of both loss functions. Once the training of DML-S2R is completed, the target value estimation of a new sample is achieved based on f^d as follows:

$$\boldsymbol{y}^* = \frac{1}{N} \sum_{i=1}^{N} \frac{f^d(\boldsymbol{x}^*, \boldsymbol{x}^l_i) - f^d(\boldsymbol{x}^l_i, \boldsymbol{x}^*)}{2} + \boldsymbol{y}^l_i. \tag{4}$$

3. EXPERIMENTAL RESULTS

Experiments were conducted on Boston Housing [12], Superconductivity [13] and Air Quality [14] datasets associated to the regression problems of housing value estimation, critical temperature estimation of a superconductor and benzene estimation for pollution monitoring, respectively. The Boston Housing dataset includes 506 samples, each of which is associated with 13 variables. We randomly selected 200 samples to construct the unlabeled sample set. The Superconductivity dataset includes 21263 samples, while each sample is associated with 81 variables. 1000 samples were randomly chosen to construct the unlabeled sample set. The Air Quality dataset consists of 9358 samples, each of which is associated with 14 variables. We randomly selected 1000 samples for the construction of the unlabeled set. In the experiments, the number of labeled samples is varied as $|S| = 10, 20, 50$ for all datasets. The rest of samples, which were not selected to the labeled and unlabeled sets, was used as the test set for each dataset. All the samples from three datasets were normalized by using min-max normalization. In the experiments, two hidden layers with 100 neurons in each were used for the SNN of DML-S2R. We trained our method for 30 epochs by using the Adam optimizer with the initial learning rate of 0.001. The results are provided in terms of mean absolute error (MAE). To perform the ablation study of the proposed DML-S2R method, we compared it with only using the first step of DML-S2R on the Superconductivity dataset. To assess the effectiveness of the proposed DML-S2R method, we compared it with the cotraining-style SSR algorithm (COREG) [15] and metric-based SSR (denoted as MSSR) method [8]. Table 1 shows the regression performances on the Superconductivity dataset obtained when: i) only the first step of the proposed DML-S2R method is applied; and ii)

Table 1: Mean absolute error (MAE) scores when the different steps of the proposed DML-S2R method were utilized (Superconductivity dataset).

Steps of DML-S2R		**Number of Labeled Samples**			
1st	2nd	10	20	50	100
✓	✗	23.5	22.8	18.2	17.2
✓	✓	**22.3**	**18.2**	**16.2**	**15.8**

Table 2: Mean absolute error (MAE) scores obtained by COREG, MSSR and the proposed DML-S2R method (Superconductivity dataset).

Method	**Number of Labeled Samples**		
	10	20	50
COREG [15]	23.1	20.5	17.2
MSSR [8]	22.4	21.2	18.1
DML-S2R (Ours)	**22.3**	**18.2**	**16.2**

both steps of DML-S2R are applied. By analyzing the table, one can see that utilizing both steps of DML-S2R results in significantly higher accuracies than using only the first step of DML-S2R with different numbers of labeled samples. As an example, DML-S2R achieves a MAE of 18.2 with 20 labeled samples, whereas using only the first step of DML-S2R results in a MAE of 22.8 with the same number labeled samples. This is due the fact that in the second step of DML-S2R, unlabeled samples are effectively exploited to learn a deep metric space that leads to more accurate target value estimation. This also shows the effectiveness of alternate learning strategy for learning both steps. Tables 2-4 shows the regression performances of COREG, MSSR and proposed DML-S2R with different number of labeled samples for all the datasets. One can see from the tables that our method provides the highest accuracies for all datasets. As an example, DML-S2R achieves a MAE of 3.3 with 50 labeled samples, whereas COREG provides a MAE of 9.5 and MSSR achieves a MAE of 12.7 with the same number of labeled samples for the Superconductivity dataset. This is due to the fact that our method effectively models the pairwise similarity of samples by using scarce labeled data, while accurately learning a deep metric space by exploiting labeled and unlabeled samples together.

4. CONCLUSION

In this paper, we have presented a novel deep metric learning-based semi-supervised regression (DML-S2R) method for parameter estimation problems. DML-S2R includes two consecutive steps. The first step aims at modelling the pairwise similarities of samples when the labeled data is scarce. This is achieved by learning to estimate the target value differences of the labeled sample pairs with a SNN. Due to this step, the proposed DML-S2R method overcomes the challenges of defining pairwise similar/dissimilar samples based on the target values of labeled samples in the framework of regression problems. The second step aims at learning a metric space that utilizes not only labeled samples but also unlabeled samples to further enrich modelling sample similarities with abundant unlabeled data. This is achieved by employing the SNN of the first step for triplet-based DML where positive-negative samples in each triplet are defined from unlabeled samples based on the SNN. The effectiveness of each step depends on each other. Accordingly, for the whole learning procedure of DML-S2R, we investigate an alternate learning strategy, in which the SNN is trained for both steps through consecutive training epochs. Due to this strategy, the encoded information by the SNN in each step becomes a guidance for learning the other step. Experimental results show the effectiveness of proposed DML-S2R compared to state-of-the-art SSR methods. It is worth noting that DML-S2R is independent from the type of the SNN architecture, and thus suitable to be integrated in any Siamese-based deep architecture. As a future work, we plan to apply DML-S2R to the parameter estimation problems in image domain (e.g., age prediction, emotion recognition, etc.).

Table 3: Mean absolute error (MAE) scores obtained by COREG, MSSR and the proposed DML-S2R method (Boston Housing dataset).

Method	**Number of Labeled Samples**		
	10	20	50
COREG [15]	8.2	7.9	5.1
MSSR [8]	6.5	5.6	4.9
DML-S2R (Ours)	**6.1**	**5.4**	**4.5**

Table 4: Mean absolute error (MAE) scores obtained by COREG, MSSR and the proposed DML-S2R method (Air Quality dataset).

Method	**Number of Labeled Samples**		
	10	20	50
COREG [15]	12.4	11.9	9.5
MSSR [8]	17.9	17.3	12.7
DML-S2R (Ours)	**10.9**	**6.0**	**3.3**

5. ACKNOWLEDGMENTS

This work is funded by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764.

6. REFERENCES

[1] B. Demir and L. Bruzzone, "A multiple criteria active learning method for support vector regression," *Pattern Recognition*, vol. 47, pp. 2558–2567, 2014.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

[4] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3476–3483, 2013.

[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Annual Conference on Computational Learning Theory*, pp. 92–100, 1998.

[6] P. Kang, D. Kim, and S. Cho, "Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing," *Expert Systems with Applications*, vol. 51, pp. 85–106, 2016.

[7] M. Timilsina, A. Figueroa, M. d'Aquin, and H. Yang, "Semi-supervised regression using diffusion on graphs," *Applied Soft Computing*, vol. 104, pp. 107188, 2021.

[8] C. Liu and Q.-H. Chen, "Metric-based semi-supervised regression," *IEEE Access*, vol. 8, pp. 30001–30011, 2020.

[9] D. Y. Choi and B. C. Song, "Semi-supervised learning for continuous emotion recognition based on metric learning," *IEEE Access*, vol. 8, pp. 113443–113455, 2020.

[10] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2283–2292, 2019.

[11] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5202–5211, 2019.

[12] D. Harrison and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.

[13] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, vol. 154, pp. 346–354, 2018.

[14] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.

[15] Z.-H. Zhou and M. Li, "Semisupervised regression with cotraining-style algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 11, pp. 1479–1493, 2007.