

# Improving Nearest Neighbor Based Multi-target Prediction Through Metric Learning

Hector Gonzalez<sup>1</sup>, Carlos Morell<sup>2</sup>, and Francesc J. Ferri<sup>3</sup>

<sup>1</sup> Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba  
hglez@uci.cu

<sup>2</sup> Universidad Central Marta Abreu (UCLV), Villa Clara, Cuba  
cmorellp@uclv.edu.cu

<sup>3</sup> Dept. Informàtica, Universitat de València, Valencia, Spain  
Francesc.Ferri@uv.es

**Abstract.** The purpose of this work is to learn specific distance functions to be applied for multi-target regression problems using nearest neighbors. The idea of preserving the order relation between input and output vectors considering their corresponding distances is used along a maximal margin criterion to formulate a specific metric learning problem. Extensive experiments and the corresponding discussion try to put forward the advantages of the proposed algorithm that can be considered as a generalization of previously proposed approaches. Preliminary results suggest that this line of work can lead to very competitive algorithms with convenient properties.

## 1 Introduction

Typical problems in pattern recognition and machine learning deal with predictors for a single discrete label or continuous value depending on whether we are dealing with classification or regression. The natural (and most common) extension to formulate the problem of predicting multiple labels/values consists of considering it as an appropriate group of independent predictors. But this approach is prone to obviate correlations among output values which may be of capital importance in many challenging and recent application domains. These methods have been coined with different names as multi-target, multi-variate or multi-response regression [2]. When the different output values are organized using more complex structures as strings or trees we talk about Structured Predictors [1, 15]. Among domain applications considered we have ecological modelling [11], gas tank control [8], remote sensing [22] and signal processing [6].

Particular methods for multi-target regression can be categorized either as problem transformation methods (when the original problem is transformed into one or several independent single-output problems), or algorithm adaptation methods (when a particular learning strategy is adapted to deal with multiple

---

F.J. Ferri—This work has been partially funded by FEDER and Spanish MEC through project TIN2014-59641-C2-1-P.

interdependent outputs). The latter methods are usually considered as more challenging as an appropriate and interpretable model is obtained usually as a subproduct for the prediction problem [2].

The purpose of this work is to improve previous approaches for multi-target regression by introducing metric learning [12] in the context of nearest neighbor methods [15]. In particular, an input-output homogeneity criterion is introduced to learn a particular distance that consistently leads to improvements according to the empirical validation carried out. In the next section, the proposed methodology is put in the context of distance based multi-target regression while Sect. 3 contains the proposal itself. The empirical section follows with details and results obtained and a final section with conclusions and further work closes the present paper.

## 2 General Notation and State of the Art

Let  $\mathbf{x} = [x_1, \dots, x_p] \in \mathbb{R}^p$ ,  $\mathbf{y} = [y_1, \dots, y_q] \in \mathbb{R}^q$ , be two random input and output vectors, respectively. Each training instance is written as  $(\mathbf{x}^j, \mathbf{y}^j) \in \mathbb{R}^p \times \mathbb{R}^q$ , and the corresponding multi-target regression problem consists of estimating a unique predictor  $h : \mathbb{R}^p \rightarrow \mathbb{R}^q$  in such a way that the expected deviation between true and predicted outputs is minimized for all possible inputs.

The most straightforward approach consists of obtaining a univariate predictor for each one of the output variables in an independent way using any of the available methods for single-target prediction [2] which constitutes the simplest of the so-called problem transformation (also known as local) methods that consist of transforming the given multi-target prediction problem into one or more single-target ones [16, 21].

The alternative approach to tackle multi-target prediction is through algorithm adaptation (also known as global) methods [2] which consist of adapting any previous strategy to deal directly with multiple targets. Global methods are interesting because they focus on explicitly capturing all interdependencies and internal relationships among targets. According to [2], these methods can be categorized as statistical, support-vector, kernel, trees or rule based, respectively. Apart from these, other strategies can be used. This is the case of one of the best known and used nonparametric methods in classification and estimation: the Nearest Neighbor (NN) family of rules [3]. Using NN for classification and estimation leads to interesting benefits as they behave quite smoothly across a wide range of applications. These methods are known to approach an optimal behavior regardless of the distance used as the number of samples grows. But nevertheless, distance becomes of capital importance in the finite case.

The  $K$ -NN for Structured Predictions (KNN-SP) method [15] has been proposed for different kind of prediction problems and for multi-target regression in particular. Using the size of the neighborhood,  $K$ , as a parameter, the KNN-SP method starts by selecting the  $K$  nearest neighbors for a given query point according to a fixed distance (usually a weighted version of the Euclidean distance).

The final prediction is constructed as the (weighted) average of the corresponding  $K$  target values. These weights are set according to the (Euclidean) distance in the target space [15]. Even though the KNN-SP is very straightforward compared to other approaches, the empirical results show that it is very competitive compared to other methods which constitute the state of the art. Moreover, neighborhood size is the only parameter to tune.

### 3 Distance Metric Learning for Multi-target Prediction (DMLMTP)

Nearest Neighbor methods have been very widely used, specially for classification. Even though it was introduced very early [18], Distance Metric Learning (DML) has been recently deeply studied as a very convenient way to improve the behavior of distance-based methods [12]. Many powerful methods have been proposed to look for the best distance (in the input space) one may have for a particular problem.

A possible way to improve the results obtained by KNN-SP is by adapting the input space distance to the particular problem according to the final goal in the same way that it has been used for classification.

Many different criteria and approaches have been proposed to learn distances for classification but all of them share the same rationale: a distance is good if it keeps same-class points close and puts points from other classes far away. Many recent approaches implement this rationale as constraints relating pairs or triplets of training points. In the case of pairs, one must select pairs of points that need to be kept close (similar points) or far away (dissimilar points). In the case of triplets, one must select some triplets,  $(\mathbf{x}^i, \mathbf{x}^j, \mathbf{x}^\ell)$ , where  $\mathbf{x}^i$  and  $\mathbf{x}^j$  are similar and should be kept close, and  $\mathbf{x}^i$  and  $\mathbf{x}^\ell$  are dissimilar and should be taken farther.

In contrast to classification problems, it is far from obvious that similar ideas are to be useful in regression problems without introducing more information about both input and output spaces. In the present work, a first attempt to learn an input distance for multi-target regression is proposed by introducing an homogeneity criterion between input and output spaces using triplets. In particular, we propose to select the same kind of triplets as in classification problems and use a different criterion for similarity. Instead of using labels, similarity between points will be established according to their outputs in such a way that the relative ordering introduced by distances in input and output spaces are preserved.

We formulate an optimization problem to learn an input distance for multi-target regression by following an approach similar to the one in [17] and also in [13, 26]. The goal is to obtain a Mahalanobis-like distance, parametrized by a matrix,  $W$ , which maximizes a margin criterion. As usual, this problem is

converted into minimizing a regularizer for  $W$  (its Frobenius norm) subject to several (soft) constraints using triplets. In our particular case we have

$$\begin{aligned} \min_{W, \rho, \xi_{ij\ell}} \quad & \frac{1}{2} \|W\|_F - \rho + \frac{1}{\nu |\mathcal{T}_K|} \sum_{i, j, \ell \in \mathcal{T}_K} \xi_{ij\ell} \\ \text{s.t.} \quad & d_W^2(\mathbf{x}^i, \mathbf{x}^\ell) - d_W^2(\mathbf{x}^i, \mathbf{x}^j) \geq \rho - \xi_{ij\ell}, \\ & \xi_{ij\ell} \geq 0, \quad \forall i, j, \ell \in \mathcal{T}_K \end{aligned}$$

where  $d_W^2(\mathbf{x}^i, \mathbf{x}^j) = (\mathbf{x}^i - \mathbf{x}^j)^T W (\mathbf{x}^i - \mathbf{x}^j)$  is the (squared) distance in the input space and the set of triplets is defined as

$$\mathcal{T}_K = \{(\mathbf{x}^i, \mathbf{x}^j, \mathbf{x}^\ell) : \mathbf{x}^j, \mathbf{x}^\ell \in \mathcal{N}_K(\mathbf{x}^i) \text{ and } d(\mathbf{y}^i, \mathbf{y}^\ell) - d(\mathbf{y}^i, \mathbf{y}^j) \geq 0\}$$

where  $\mathcal{N}_K(\mathbf{x})$  is the considered neighborhood around  $\mathbf{x}$ .

Note that the formulation of the optimization problem is the same used for other metric learning and support vector learning approaches and the main change is in the way the particular restrictions have been selected.

In the formulation above, we must introduce an extra constraint to make the matrix  $W$  positive semi-definite. This makes the problem considerably more difficult but there are a number of ways in which this can be tackled [13, 26]. Nevertheless, in this preliminary work we will simplify the above formulation further. On one hand, we consider only a diagonal matrix,  $W = \mathbf{w} = [w_1, \dots, w_p]$ , and on the other hand, we will introduce the corresponding restrictions,  $w_i \geq 0$ ,  $i = 1, \dots, q$  into the above optimization. The corresponding dual problem can be written in terms of two new sets of variables as

$$\begin{aligned} \min_{\alpha_i, \lambda_j} \quad & \frac{1}{2} \left( \boldsymbol{\alpha}^T H \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^T \boldsymbol{\phi} \boldsymbol{\lambda} + \boldsymbol{\lambda}^T \boldsymbol{\lambda} \right) \\ \text{s.t.} \quad & \sum_{i=1}^{|\mathcal{T}_K|} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq \frac{1}{\nu |\mathcal{T}_K|} \quad i = 1, \dots, |\mathcal{T}_K| \\ & \lambda_j \geq 0 \quad j = 1, \dots, q \end{aligned}$$

$\boldsymbol{\phi} \in \mathbb{R}^{|\mathcal{T}_K| \times q}$  is a matrix with a row,  $(\mathbf{x}^i - \mathbf{x}^\ell) \circ (\mathbf{x}^i - \mathbf{x}^\ell) - (\mathbf{x}^i - \mathbf{x}^j) \circ (\mathbf{x}^i - \mathbf{x}^j)$ , for each considered triplet where  $\circ$  is the Hadamard or entrywise vector product. The kernel matrix is  $H = \boldsymbol{\phi} \boldsymbol{\phi}^T$  and the weight vector is obtained as  $\mathbf{w} = \boldsymbol{\alpha}^T \boldsymbol{\phi} + \boldsymbol{\lambda}$ .

An adhoc solver using an adapted SMO approach [10, 14] has been implemented specifically for this work. This solver is able to arrive to relatively good results in reasonable times for all databases considered in the empirical work carried out as will be shown in the next section.

## 4 Experiments

In this section, we describe the experimental setup and discuss the main results of the proposed DMLMTP algorithm. In the first place, we present technical

**Table 1.** Datasets used in the experimentation and corresponding details. Datasets partitioned in train and test subsets are indicated by the corresponding two sizes in the second column.

Datasets	Instances	Attributes	Targets
Waterquality [5]	1060	16	14
EDM [9]	154	16	2
Solar Flare 1 [23]	323	10	3
Solar Flare 2 [23]	1066	10	3
jura [19]	359	15	3
enb [19]	768	8	2
slump [19]	103	7	3
andro [19]	49	30	6
osales [24]	639	413	12
scpf [25]	1137	23	3
atp1d [19]	201/136	411	6
atp7d [19]	188/108	411	6
rf1 [19]	4108/5017	64	8
rf2 [19]	4108/5017	576	8
OES97 [19]	334	263	16
OES10 [19]	410	298	16

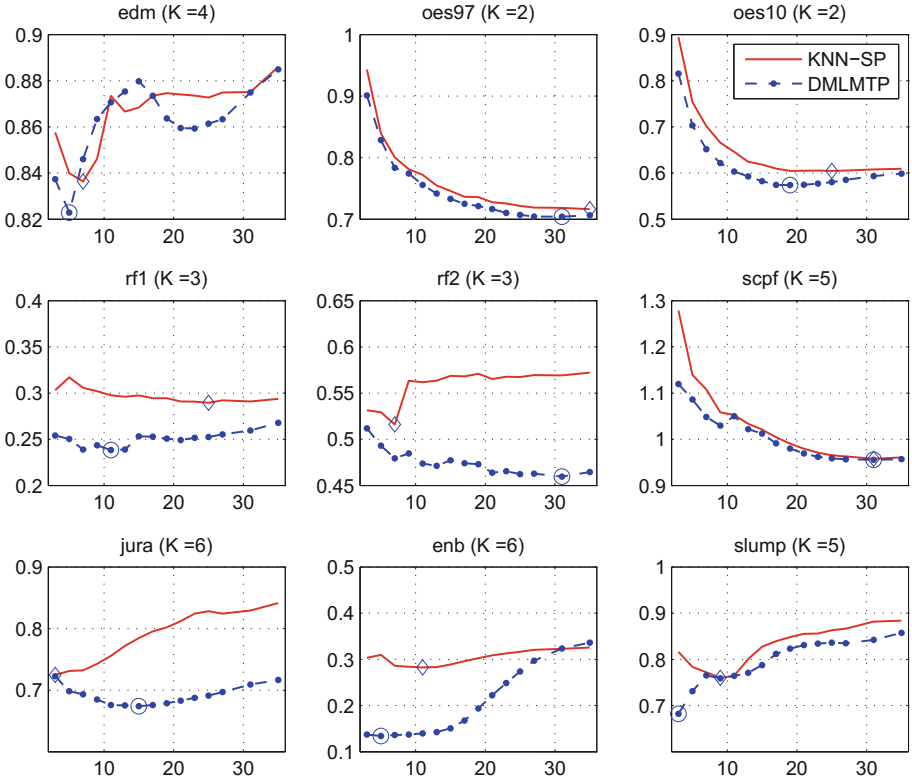
details related to the datasets, parameter setup and implementations. Next, we present comparative results when using the learned distance compared to the Euclidean one when predicting multivariate outputs with the KNN-SP approach over fifteen datasets publicly available for multi target prediction.

In the experiments we distinguish between the number of neighbors used to learn the distance using DMLMTP,  $K$ , and the number of neighbors used to obtain the final prediction using the KNN-SP approach,  $k_p$ . The value of  $K$  should be small to keep the number of triplets small for efficiency reasons. For all experiments reported in this paper, the number of nearest neighbors for training in DMLMTP was set to  $K = 2, \dots, 6$  while the neighborhood sizes for prediction have been taken as odd values from 3 to 35. The final prediction is done computing the average of the target values of these  $k_p$  nearest neighbors.

In the experiments, 5-fold cross-validation has been used on each dataset except for 4 of them that have been split into train and test subsets for efficiency and compatibility reasons. Table 1 summarizes the main details [2, 21]. The cross validation procedure has been integrated into MULAN software package [20].

As in other similar works, we use the average Relative Root Mean Squared Error (aRRMSE) given a test set,  $D_{test}$ , and a predictor,  $h$ , which is given as

$$aRRMSE(h; D_{test}) = \frac{1}{q} \sum_{i=1}^q \sqrt{\frac{\sum_{(\mathbf{x}, \mathbf{y}) \in D_{test}} (\hat{y}_i - y_i)^2}{\sum_{(\mathbf{x}, \mathbf{y}) \in D_{test}} (\bar{y}_i - y_i)^2}} \quad (1)$$



**Fig. 1.** aRRMSE values corresponding to different neighborhood sizes,  $k_p$ , using Euclidean (KNN-SP) and learned (DMLMTP) distances. The best neighborhood size used by DMLMTP is indicated along the name of each database.

where  $\bar{y}$  is the mean value of the target variable  $\mathbf{y}$ , and  $\hat{y} = h(\mathbf{x})$ . We use the Wilcoxon signed rank test and the Friedman procedure with different post-hoc tests to compare algorithms over multiple datasets [4, 7].

Figure 1 contains the aRRMSE versus the neighborhood size,  $k_p$  for 9 datasets out of the 16 considered. Only the best neighborhood size used for training,  $K$  is shown. Moreover, the best results in the curves are marked with a circle and a diamond, respectively. These best results are shown for all the datasets in Table 2.

Contrary to our expectations, the best performance for DMLMTP over large datasets is obtained with small  $K$ . This could be strongly related to the growth in the number of triplets that violate the considered constraints.

The last columns in Table 2 contain the absolute difference between aRRMSE for KNN-SP and DMLMTP, its sign and the average ranking with regard to absolute differences. The DMLMTP method is better with a significance level of 5% according to the Wilcoxon test that leads to a p-value of 6.1035e-5. For all

**Table 2.** aRRMSE obtained for DMLMTP and KNN-SP algorithms on each dataset along with comparison details.

Dataset	DMLMTP	KNN-SP	abs	sgn	$R_i$
sf1	0.971	0.973	0.002	+	2.5
sf2	0.976	0.977	0.001	+	1.5
wq	0.947	0.948	0.001	+	1.5
edm	0.823	0.836	0.013	+	6
oes97	0.704	0.716	0.012	+	5
oes10	0.574	0.605	0.031	+	7
atp1d	0.415	0.451	0.036	+	8
atp7d	0.515	0.605	0.090	+	13
rf1	0.238	0.290	0.052	+	10
osales	1.012	1.012	0		-
scpf	0.956	0.958	0.002	+	2.5
jura	0.674	0.725	0.051	+	9
enb	0.134	0.283	0.149	+	15
slump	0.682	0.760	0.078	+	12
andro	0.799	0.931	0.132	+	14
rf2	0.460	0.516	0.056	+	11

datasets, DMLMTP has equal or better performance than (Euclidean) KNN-SP and the difference increases for datasets of higher dimensionality. This situation could be related to the learned input transformation that generates some values equal to zero and ignores some irrelevant attributes. In fact, if we compute the sparsity index of the corresponding transformation vector, as the relative number of zeros with regard to dimensionality, we obtain for our algorithm values below 0.5 except for datasets osales, rf1 and scpf.

## 5 Concluding Remarks and Further Work

An attempt to improve nearest neighbor based multi-target prediction has been done by introducing an specific distance metric learning algorithm. The mixing of these strategies has lead to very competitive results in the preliminary experimentation carried out. In a wide range of situations and for large variations of the corresponding parameters, the proposal behaves smoothly over the datasets considered paving the way to develop more specialized algorithms. Future work is being planned in several directions. On one hand, different optimization schemes can be adopted both to improve efficiency and performance. On the other hand, different formulations can be adopted by establishing more accurate constraints able to properly capture all kinds of dependencies among input and output vectors in challenging multi output regression problems.

## References

1. Bakir, G.H., Hofmann, T., Schölkopf, B., Smola, A.J., Taskar, B., Vishwanathan, S.V.N.: Predicting Structured Data (Neural Information Processing). The MIT Press, Cambridge (2007)
2. Borchani, H., Varando, G., Bielza, C., Larrañaga, P.: A survey on multi-output regression. *Wiley Interdisc. Rev. Data Mining Knowl. Discov.* **5**(5), 216–233 (2015)
3. Dasarathy, B.V.: Nearest Neighbor (NN) Norms: NN Pattern Classification. IEEE Computer Society, Washington (1990)
4. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006)
5. Džeroski, S., Demšar, D., Grbović, J.: Predicting chemical parameters of river water quality from bioindicator data. *Appl. Intell.* **13**(1), 7–17 (2000)
6. Fernández, M.S., de Prado-Cumplido, M., Arenas-García, J., Pérez-Cruz, F.: SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Trans. Sig. Process.* **52**(8), 2298–2307 (2004)
7. García, S., Fernández, A., Luengo, J., Herrera, F.: A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft. Comput.* **13**(10), 959–977 (2009)
8. Han, Z., Liu, Y., Zhao, J., Wang, W.: Real time prediction for converter gas tank levels based on multi-output least square support vector regressor. *Control Eng. Pract.* **20**(12), 1400–1409 (2012)
9. Karalič, A., Bratko, I.: First order regression. *Mach. Learn.* **26**(2–3), 147–176 (1997)
10. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt’s SMO algorithm for SVM classifier design. *Neural Comput.* **13**(3), 637–649 (2001)
11. Kocev, D., Džeroski, S., White, M.D., Newell, G.R., Griffioen, P.: Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecol. Model.* **220**(8), 1159–1168 (2009)
12. Kulis, B.: Metric learning: a survey. *Found. Trends Mach. Learn.* **5**(4), 287–364 (2012)
13. Perez-Suay, A., Ferri, F.J., Arevalillo, M., Albert, J.V.: Comparative evaluation of batch and online distance metric learning approaches based on margin maximization. In: *IEEE International Conference on Systems, Man, and Cybernetics, Manchester, SMC 2013, UK*, pp. 3511–3515 (2013)
14. Platt, J., et al.: Fast training of support vector machines using sequential minimal optimization. In: *Advances in Kernel Methods—Support Vector Learning*, vol. 3 (1999)
15. Pugalj, M., Džeroski, S.: Predicting structured outputs  $k$ -nearest neighbours method. In: Elomaa, T., Hollmén, J., Mannila, H. (eds.) *DS 2011. LNCS (LNAI)*, vol. 6926, pp. 262–276. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-24477-3\\_22](https://doi.org/10.1007/978-3-642-24477-3_22)
16. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. *Mach. Learn.* **85**, 333–359 (2011)
17. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: *Advances in Neural Information Processing Systems (NIPS)*, p. 41 (2004)
18. Short, R., Fukunaga, K.: The optimal distance measure for nearest neighbor classification. *IEEE Trans. Inf. Theory* **27**(5), 622–627 (1981)



19. Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., Vlahavas, I.: Multi-target regression via input space expansion: treating targets as inputs. *Mach. Learn.* **104**(1), 55–98 (2016)
20. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: a Java library for multi-label learning. *J. Mach. Learn. Res.* **12**, 2411–2414 (2011)
21. Tsoumakas, G., Spyromitros-Xioufis, E., Vrekou, A., Vlahavas, I.: Multi-target regression via random linear target combinations. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) *ECML PKDD 2014. LNCS (LNAI)*, vol. 8726, pp. 225–240. Springer, Heidelberg (2014). doi:[10.1007/978-3-662-44845-8\\_15](https://doi.org/10.1007/978-3-662-44845-8_15)
22. Tuia, D., Verrelst, J., Alonso-Chorda, L., Pérez-Cruz, F., Camps-Valls, G.: Multi-output support vector regression for remote sensing biophysical parameter estimation. *IEEE Geosci. Remote Sens. Lett.* **8**(4), 804–808 (2011)
23. <http://archive.ics.uci.edu/ml/datasets/Solar+Flare>
24. <https://www.kaggle.com/c/online-sales>
25. <https://www.kaggle.com/c/see-click-predict-fix>
26. Wang, F., Zuo, W., Zhang, L., Meng, D., Zhang, D.: A kernel classification framework for metric learning. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(9), 1950–1962 (2015)