Schedae Informaticae Vol. 24 (2015): 83–92
doi: 10.4467/20838476SI.15.008.3030

Mixture of Metrics Optimization for Machine Learning Problems

Magdalena Wiercioch, Marek Śmieja
Faculty of Mathematics and Computer Science
ul. Łojasiewicza 6, 30-348 Kraków
e-mail: {*magdalena.wiercioch, marek.smieja*}*@ii.uj.edu.pl*

Abstract. The selection of data representation and metric for a given data set is one of the most crucial problems in machine learning since it affects the results of classification and clustering methods. In this paper we investigate how to combine a various data representations and metrics into a single function which better reflects the relationships between data set elements than a single representation-metric pair. Our approach relies on optimizing a linear combination of selected distance measures with use of least square approximation. The application of our method for classification and clustering of chemical compounds seems to increase the accuracy of these methods.

Keywords: metric learning, classification, clustering, chemical compound activity, fingerprint.

1. Introduction

The appropriate choice of data representation and metric are one of the main problems in machine learning tasks [7, 20, 26]. Their selection affects directly clustering results and classification methods. In this paper we show how to combine various representations and metrics into one function which better reflects the geometry of the underlying space.

In general, metric learning aims at adapting some pairwise real-valued function to the specific problem making use of the information retrieved from the process of training. A lot of metric learning techniques were proposed to select or optimize a metric for particular data sets. This includes algorithms such as Locally Linear

Embedding [17], Multidimensional Scaling [5]. In 2002 Xing et al. explored learning a Mahalanobis metric [24]. Finally, Taketa et al. published an algorithm for kernel regression [23].

Our motivation was to design a method which optimizes existing metrics and representations without constructing of completely new distance measure. Such an approach can be useful when a lot of distance measures are available but none of them give a satisfactory results. This situation appears very often in practice e.g., in chemoinformatics where one would like to detect the chemical compounds active on particular diseases with use of computer methods only [8]. This is a very important problem since the appropriate classification of compounds in terms of their activities enables to decrease extensively the computational time and costs of finding new drugs. A lot of compounds representations and metrics are available but none reflects the activity satisfactory.

We assume that a real distance between data set elements x, y is described by an unknown real valued function $a(x, y)$ (it can be the difference between the activity levels of two chemical compounds which cannot be calculated analytically but has to be measured in the experiment). We propose to build a linear combination of existing distance measures and optimize its coefficients to approximate values of the function a. In particular, given distance measures $d_1, \ldots, d_n$, we construct a linear regression model:

$$a(x, y) \approx w_1 d_1(x, y) + \ldots + w_n d_n(x, y)$$

and calculate coefficients $w_1, \ldots, w_n$ with use of least square estimation[1].

The application of introduced method is presented on various data sets of chemical compounds (six biological receptor ligands are considered). It is shown that constructed function provides better classification and clustering results than the use of any individual distance measure.

The paper is organized as follows. The next section gives a background of proposed method of the mixture of metrics optimization. Then, it provides basic information about chemical compounds representations and activity. In the third section we examine the capabilities of our algorithm for real data. Finally, the conclusion is given.

2. Mixture of distance measures

We begin this section with a formulation of our optimization problem. Then, we use a least square estimator to solve this approximation. We will later place our metod in the context of chemical compounds.

Optimization problem. Let X be a data set. We assume that the dissimilarity between elements of X is described by an unknown function $a : X \times X \to [0, \infty)$. The value $a(x, y)$, for $x, y \in X$, might be a measurement performed in the experiment or it can be assigned by an external system. For instance, in the case of chemical compounds it can be a difference between compounds activities or an expert indication

[1] In fact we consider a dissimilarity measure $w_0 + w_1\, d_1 + \ldots w_n\, d_n$ rather than a metric.

if two compounds belong to the same chemical group. The goal is to find a function $d : X \times X \to \mathbb{R}$ which can be easily calculated analytically and such that $d(x, y)$ approximates the value $a(x, y)$, for $x, y \in X$.

Our approach relies on an appropriate mixing of various available distance measures. Given n distance measures $d_1, \ldots, d_n$ we look for numbers $w_1, \ldots, w_n \geq 0$ such that

$$\widetilde{d_w}(x, y) := w_1 d_1(x, y) + \ldots + w_n d_n(x, y)$$

approximates $a(x, y)$, for $x, y \in X$.

Observation 1 *If $d_1, \ldots, d_n$ define metrics in X and $w_1, \ldots, w_n \geq 0$ then $\widetilde{d_w}$ is also a metric in X.*

The structure of predicted function a can be very complex. Therefore, for practical reasons we are rather interested in finding a dissimilarity measure (not necessarily a metric) of the form:

$$d_w(x, y) := w_0 + w_1 d_1(x, y) + \ldots + w_n d_n(x, y),$$

where $w_0, w_1, \ldots, w_n \in \mathbb{R}$ and $x, y \in X$.

A lot of criteria can be used to measure the discrepancy between d_w and a. In this paper we assume a linear regression model and use a least square estimator to minimize the square error, i.e.

$$\sum_{x,y \in X} (a(x, y) - d_w(x, y))^2.$$

Lest square estimator. For a convenience of the reader let us briefly recall a linear regression model and a least square estimator. Let

$$\mathbf{D} = \begin{pmatrix} 1 & d_1(x_1, y_1) & \cdots & d_n(x_1, y_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d_1(x_k, y_k) & \cdots & d_n(x_k, y_k) \end{pmatrix}$$

be the $(n + 1) \times k$ dimensional matrix of observations and let

$$\mathbf{a} = \begin{pmatrix} a(x_1, y_1) \\ \vdots \\ a(x_k, y_k) \end{pmatrix}$$

be the vector with corresponding values of function a. We assume that there exists a linear dependence between $\mathbf{a}$ and $\mathbf{D}$, i.e. there exists a real valued vector

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix}$$

such that

$$\mathbf{a} = \mathbf{D}\mathbf{w} + \varepsilon, \tag{1}$$

where

$$\varepsilon = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

is an unobserved random noise.

The following theorem gives a form and the properties of least square estimator of vector $\mathbf{w}$.

Theorem 1 (Gauss-Markov Theorem [1]) *If errors ε_i have expectations zero, are uncorrelated and have equal (bonded) variance then the least squares estimator*

$$\hat{\mathbf{w}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{a}$$

is the minimum variance linear unbiased estimator of $\mathbf{w}$.

Preliminaries of chemical compounds. The main goal of chemoinformatics is to support the process of new drug candidates finding. One of the tasks is to find compounds which are active on particular biological targets. Therefore, the clustering and classification of chemical compounds with respect to their chemical activity is important.

Chemical compounds are usually represented by fingerprints, i.e. high dimensional binary strings where a given bit indicates the presence or absence of particular feature of compound. Since a lot of features can be taken into account, various fingerprint representations were constructed. The length of popular representations varies from 166 (MACCS) to 4860 bits (Klekota-Roth [14]). Although the fingerprint representations can be very long, they do not provide the uniqueness. In the other words, in every representation there exist chemical compounds which have identical fingerprints. Clearly the concatenation of the fingerprints allows for a better (but not ideal) distinction of compounds [6].

One can determine the biological activity of a compound by examining a binding constant[2] K_i measured in nanomols (nM) [25]. The prediction of compound's activity is usually repeated several times and then averaged. The possible deviances might be due to many conditions (e.g., temperature, pressure). What is more, the border between active and inactive compounds is not established. In practice, for internal research some assumptions have to be made. For example, the compound may be considered as active if $K_i \leq 100$ while for $K_i \geq 1000$ the compound is considered as inactive (other compounds are usually not taken into account).

We also follow this approach in clustering and classification processes. However, since the regression model assumes a random noise, the values of K_i can be used to define the dependent variable. We want to design a dissimilarity measure which gives low values for compounds with similar activities while high values are assigned for compounds with different values of K_i [3]. We define a dependent variable as a difference between activities of two compounds while as the explanatory variables we

[2] To obtain more reliable information about activity one can also use IC_{50}, EC_{50}, K_d values.

[3] Roughly speaking, the constructed measure will indicate that two compounds are similar if their activity levels are close.

Table 1. Overview of considered data sets. Table contains the names and roles of used receptors ligands and the number of active and inactive compounds included in each dataset.

Receptor name	Role	Actives	Inactives
M_1	modulates few of physiological functions	759	938
h_1	has an impact on pathophysiological conditions	635	545
5-HT_7	influences on various neurological processes, such as aggression	704	339
5-HT_{2A}	has an impact on central nervous system	1835	851
5-HT_6	mediates both excitatory and inhibitory neurotransmission	1490	341
5-HT_{2C}	has an impact on central nervous system	1210	926

assume the distances between two compounds with respect to given metrics and representations. Thus, the regression model can be written as:

$$|K_i(x) - K_i(y)| = w_0 + w_1\, d(x,y) + \ldots + w_n\, d(x,y) + \varepsilon,$$

for $x, y \in X$.

3. Experiments

We have examined our method on six data sets of chemical compounds, each representing one receptor ligands, Table 1. The dissimilarity measures were learned and then tested either in classification or in clustering processes assuming 5-fold cross validation [15]. The classification and clustering results were evaluated with use of adjusted Rand index (ARI) [11] which is a well-known measure of agreement between two partitions. ARI assumes its minimum of 0 in the case of completely independent partitions while for ideal agreement it gives value of 1.

In the first experiment we combined four fingerprint representations (Klekota Roth, Extended, Substructure and PubChem) with two dissimilarity measures (Buser and Tanimoto) [21] to define the explanatory variables. Optimized metrics were assessed in the k-NN classification [4] as well as in k-means [16] and hierarchical clusterings [22]. The results were compared with analogical classifications and clusterings obtained with use of every single representation-metric pair, Table 2.

In all cases the optimized dissimilarity measures give better results than a single representation-metric pair. The increase in precision is especially visible in the case of classification (about 10%). The highest value of ARI has been obtained for 5-HT_{2A} receptor (ARI = 0.7). Moreover, the results show that none of the representation-metric

Table 2. Classification and clustering results measured by ARI for optimized dissimilarity measures compared with effects obtained with use of single representation-metric pairs. The names of the last six columns indicate applied representation-metric pairs and have the following meanings: Buser-Klekota Roth (B-KR), Buser-Extended (B-Ext), Buser-Substructure (B-Subs), Tanimoto-Klekota Roth (T-KR), Tanimoto-Extended (T-Ext), Tanimoto-Substructure (T-Subs).

(a) k-NN classification

Receptor	Optimized	B-KR	B-Ext	B-Subs	T-KR	T-Ext	T-Subs
M_1	**0.67**	0.57	0.55	0.57	**0.58**	0.54	0.54
h_1	**0.65**	0.59	0.56	0.52	0.58	**0.6**	0.57
5-HT_7	**0.69**	**0.63**	0.61	0.56	0.58	0.59	0.56
5-HT_6	**0.68**	0.6	**0.62**	0.6	0.57	0.57	0.57
5-HT_{2C}	**0.66**	0.61	0.59	0.49	**0.63**	0.56	0.5
5-HT_{2A}	**0.7**	**0.64**	0.61	0.59	0.64	0.59	0.54

(b) k-means clustering

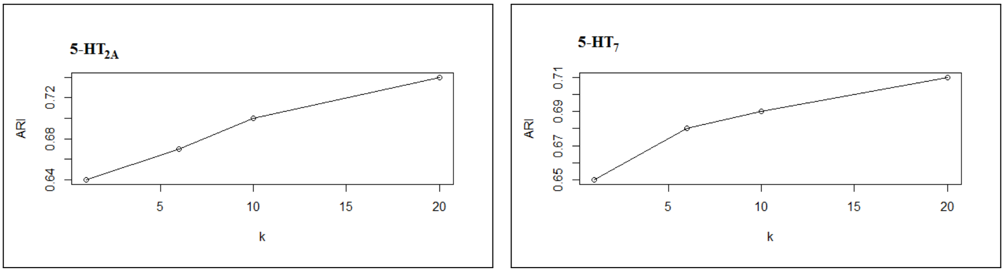
Receptor name	Optimized	B-KR	B-Ext	B-Subs	T-KR	T-Ext	T-Subs
M_1	**0.4**	**0.39**	0.36	0.37	0.36	0.37	0.34
h_1	**0.3**	**0.28**	0.27	0.24	0.26	0.26	0.27
5-HT_7	**0.52**	0.48	**0.49**	0.46	0.48	0.45	0.45
5-HT_6	**0.33**	0.3	0.3	**0.31**	0.31	0.29	0.27
5-HT_{2C}	**0.46**	**0.44**	0.43	0.4	0.42	0.39	0.39
5-HT_{2A}	**0.35**	**0.31**	0.3	0.31	0.3	0.31	0.28

(c) Hierarchical clustering

Receptor name	Optimized	B-KR	B-Ext	B-Subs	T-KR	T-Ext	T-Subs
M_1	**0.45**	0.4	**0.41**	0.35	0.39	0.37	0.36
h_1	**0.23**	**0.19**	0.15	0.17	0.19	0.17	0.16
5-HT_7	**0.41**	0.35	0.33	0.35	**0.36**	0.34	0.33
5-HT_6	**0.4**	0.36	**0.37**	0.35	0.37	0.34	0.34
5-HT_{2C}	**0.52**	**0.48**	0.46	0.45	0.46	0.44	0.45
5-HT_{2A}	**0.42**	0.35	0.33	0.34	**0.36**	0.34	0.32

pair gives the highest agreement for all receptors. In our approach the dissimilarity measure is optimized on a collection of metrics automatically and the user does not have to specify one representation-metric pair used for calculation. This makes our method robust to the choice of distance measure.

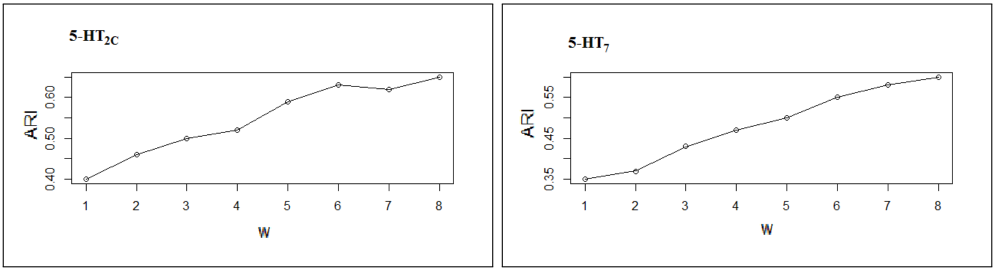
Inspired by above-mentioned results, we investigated the influence of neighbors number on classification results. Four settings were considered: 1, 6, 10 and 20. One can observe that the classification results were even improved when more neighbors have been taken into account (Figure 1a and 1b).



(a) 5-HT$_{2A}$ **(b)** 5-HT$_7$

Figure 1. k-NN results measured by ARI after optimization for different numbers k.

In order to explore how the results vary when the number of explanatory variables in regression model changes, we focused on two receptors: 5-HT$_{2C}$ and 5-HT$_7$. The following procedure was considered. All representation-metric pairs were ordered with respect to the highest ARI obtained in k-NN classification. Then, several regression models were built. The first one included only the best representation-metric pair. The subsequent models were constructed by adding one more explanatory variable to the model (with respect to the highest ARI values). The results shown in Figures 2a and 2b indicate the gradual increase in classification results.



(a) 5-HT$_{2C}$ **(b)** 5-HT$_7$

Figure 2. The more explanatory variables in model, the higher values ARI yields. The following metric-representation pairs were considered: Buser-Klekota Roth, Buser-Extended, Buser-Substructure, Buser-PubChem, Tanimoto-Klekota Roth, Tanimoto-Extended, Tanimoto-Substructure, Tanimoto-PubChem.

4. Conclusion

In this paper we have addressed the problem of metric learning. According to our approach, a more relevant metric can be obtained by defining a single function which combines various data representation-metric pairs. More precisely, we have demonstrated that for real-world data (chemical compounds) such an optimized metric can be learned and improve the performance of metric-based algorithms. Taken as a whole, our results exhibit the promise and broad applicability of proposed approach in methods using metrics.

Acknowledgement

This research was partially supported by the National Centre of Science (Poland) Grant No. 2014/13/N/ST6/01832. The authors are very grateful to Professor Jacek Tabor for many useful suggestions and to Dawid Warszycki for fruitful discussions about chemical aspects of fingerprints.

5. References

[1] Aczel A., Sounderpandian J., *Complete Business Statistics.* McGraw Hill, New York 2009.

[2] Atkeson C., Moore A., Schaal S., *Locally weighted learning.* Artificial Intelligence Review, 1997, 11, pp. 11–73.

[3] Bar-Hillel A., Hertz T., Shental N., Weinshall D., *Learning a Mahalanobis metric from equivalence constraints.* Journal of Machine Learning Research, 2005, 6, pp. 937–965.

[4] Cover T., Hart P., *Nearest Neighbor Pattern Classification.* IEEE Transactions on Information Theory, 1967, 13, pp. 21–27.

[5] Cox T.F., Cox M.A.A., *Multidimensional Scaling.* Chapman and Hall, London 1994.

[6] Deng Z., Chuaqui C., Singh J., *Knowledge-based design of target-focused libraries using protein-ligand interaction constraints.* Journal of Medicinal Chemistry, 2006, 49(2), pp. 490–500.

[7] Domeniconi C., Gunopulos D., *Adaptive nearest neighbor classification using support vector machines.* Advances in Neural Information Processing Systems, 2002, 14, pp. 665–672.

[8] Geppert H., Vogt M., Bajorath J., *Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation.* Journal of Chemical Information and Modeling, 2010, 50, pp. 205–216.

[9] Goldberger J., Roweis S., Hinton G., Salakhutdinov R., *Neighbourhood Components Analysis.* Advances in Neural Information Processing Systems, 2004, 17, pp. 513–520.

[10] Hastie T., Tibshirani R., *Discriminant Adaptive Nearest Neighbor Classification.* IEEE Trans. Pattern Anal. Mach. Intell., 1996, 18, pp. 607–616.

[11] Hubert L., Arabie P., *Comparing partitions.* Journal of Classification, 1985, 2, pp. 193–218.

[12] Jaakkola T.S., Haussler D., *Exploiting Generative Models in Discriminative Classifiers.* Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, 1999, pp. 487–493.

[13] Kedem D., Tyree S., Weinberger K.Q., Sha F., Lanckriet G., *Non-linear Metric Learning.* Advances in Neural Information Processing Systems, 2012, 25, pp. 2582–2590. Available via http://books.nips.cc/papers/files/nips25/NIPS2012_1223.pdf.

[14] Klekota J., Roth F.P., *Chemical Substructures That Enrich for Biological Activity.* Bioinformatics 2008, 21, pp. 2518–2525.

[15] Kohavi R., *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95), 1995, pp. 1137–1143.

[16] Lloyd S., *Least Squares Quantization in PCM.* IEEE Trans. Inf. Theor., 1982, 28, pp. 129–137.

[17] Roweis S.T., Saul L.K., *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000, 290, pp. 2323–2326.

[18] Scholkopf B., Smola A.J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, 2001.

[19] Shalev-Shwartz S., Singer Y., Ng A.Y., *Online and Batch Learning of Pseudometrics.* Proceedings of the Twenty-first International Conference on Machine Learning (ICML '04), 2004, pp. 743–750.

[20] Shental N., Hertz T., Weinshall D., Pavel M., *Adjustment Learning and Relevant Component Analysis.* Proceedings of the 7th European Conference on Computer Vision-Part IV (ECCV '02), 2002, pp. 776–792.

[21] Śmieja M., Warszycki D., Tabor J., Bojarski A.J., *Asymmetric Clustering Index in a Case Study of 5-HT$_{1A}$ Receptor Ligands.* PloS ONE 9(7): e102069, doi:10.1371/journal.pone.0102069, 2014.

[22] Sneath P.H.A., *The Application of Computers to Taxonomy.* J. Gen. Microbiol., 1957, 17, pp. 201–226.

[23] Takeda H., Farsiu S. and Milanfar P., *Robust kernel regression for restoration and reconstruction of images from sparse noisy data.* IEEE International Conference on Image Processing, 2006, pp. 1257–1260.

[24] Xing E.P., Ng A.Y., Jordan M.I., Russell S., *Distance Metric Learning, With Application To Clustering With Side-Information,.* Advances in Neural Information Processing Systems, 2003, 15, pp. 505–512.

[25] Warszycki D., Mordalski S., Kristiansen K., Kafel R., Sylte I., Chilmonczyk, Z., Bojarski A. J., *A Linear Combination of Pharmacophore Hypotheses as a New Tool in Search of New Active Compounds An Application for 5-HT$_{1A}$ Receptor Ligands.* PloS ONE 8(12): e84510, doi:10.1371/journal.pone.0084510, 2013.

[26] Weinberger K.Q., Saul L.K., *Distance Metric Learning for Large Margin Nearest Neighbor Classification.* J. Mach. Learn. Res., 2009, 10, pp. 207–244.

[27] Weinberger K.Q., Saul L.K., *Fast solvers and efficient implementations for distance metric learning.* ACM International Conference Proceeding Series, 2008, 307, pp. 1160–1167.