Mixture of metrics optimization for machine learning problems

Magdalena Wiercioch and Marek Śmieja
Faculty of Mathematics and Computer Science, Jagiellonian University

TFML 2015
Będlewo, February 16-21

Goals of this Work

- How to select data representation and metric for a given data set?

Goals of this Work

- How to select data representation and metric for a given data set?
- Combining various data representations and metrics.

Goals of this Work

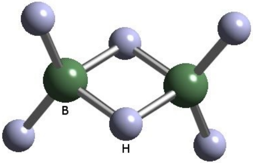
- How to select data representation and metric for a given data set?
- Combining various data representations and metrics.
- Optimizing a linear combination of selected distance measures.

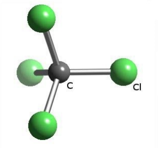
Outline

- Motivation
- Background
- Approach
- Experimental results
- Conclusion

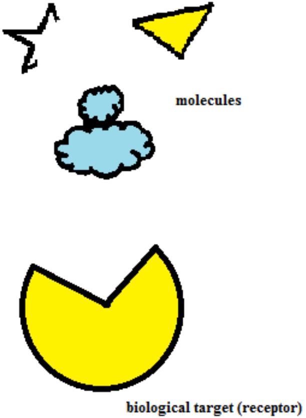
Motivation

real - life problem of chemoinformatics





Motivation



molecules
biological target (receptor)

they do not activate this receptor

Representation of molecules

Fingerprints are binary strings where a given bit indicates the absence or presence of particular pattern.

Representation of molecules

Fingerprints are binary strings where a given bit indicates the absence or presence of particular pattern. Problems:

- high dimensionality

Representation of molecules

Fingerprints are binary strings where a given bit indicates the absence or presence of particular pattern. Problems:

- high dimensionality
- they are not unique

Biological activity

- IC_{50}, EC_{50}, K_d

Biological activity

- IC_{50}, EC_{50}, K_d
- a binding constant K_i was used

Biological activity

- IC_{50}, EC_{50}, K_d
- a binding constant K_i was used
- prediction of molecule's activity is repeated several times

Biological activity

- IC_{50}, EC_{50}, K_d
- a binding constant K_i was used
- prediction of molecule's activity is repeated several times
- the chemical compound were considered as active if $K_i \leq 100$ while for $K_i \geq 1000$ - inactive

Intuitively: design a measure which gives low values for compounds with similar activities while high values are assigned for compounds with different values of K_i.

Intuitively: design a measure which gives low values for compounds with similar activities while high values are assigned for compounds with different values of K_i.

METRIC LEARNING

Background

- Multidimensional Scaling (1994)
- Locally Linear Embedding (Roweis and Saul, 2000)
- learning a Mahalanobis metric by Xing et al. (2003)
- kernel regression (Takeda et al., 2006)
- ...

Our aims

- optimize existing metrics and representations
  - use of combination distance measures
  - coefficients
- improve classification and clustering results

Our aims

- optimize existing metrics and representations
  - use of combination distance measures
  - coefficients
- improve classification and clustering results

$$a(x, y) \approx \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$$

Optimization

- X - data set
- $a : X \times X \to [0, \infty)$

Optimization

- X - data set
- $a : X \times X \to [0, \infty)$
- $d : X \times X \to \mathbb{R}$

Optimization

- X - data set
- $a : X \times X \to [0, \infty)$
- $d : X \times X \to \mathbb{R}$
- $\widetilde{d_\omega}(x, y) := \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$

Optimization

- X - data set
- $a : X \times X \to [0, \infty)$
- $d : X \times X \to \mathbb{R}$
- $\widetilde{d_\omega}(x, y) := \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$
- in practice
  $d_\omega(x, y) := \omega_0 + \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$

Optimization

- X - data set
- $a : X \times X \to [0, \infty)$
- $d : X \times X \to \mathbb{R}$
- $\widetilde{d_\omega}(x, y) := \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$
- in practice
  $d_\omega(x, y) := \omega_0 + \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$ or less formally
  $|K_i(x) - K_i(y)| = \omega_0 + \omega_1 d(x, y) + \ldots + \omega_n d(x, y) + \epsilon,$

Optimization

- X - data set
- $a : X \times X \to [0, \infty)$
- $d : X \times X \to \mathbb{R}$
- $\widetilde{d_\omega}(x, y) := \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$
- in practice
  $d_\omega(x, y) := \omega_0 + \omega_1 d_1(x, y) + \ldots + \omega_n d_n(x, y)$ or less formally
  $|K_i(x) - K_i(y)| = \omega_0 + \omega_1 d(x, y) + \ldots + \omega_n d(x, y) + \epsilon,$
- $\sum_{x,y \in X} (a(x, y) - d_\omega(x, y))^2$

Data sets

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

Dissimilarity metrics

- Buser: $\frac{cd+c}{cd+a+b-c}$
- Tanimoto: $\frac{c}{a+b-c}$

kNN

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

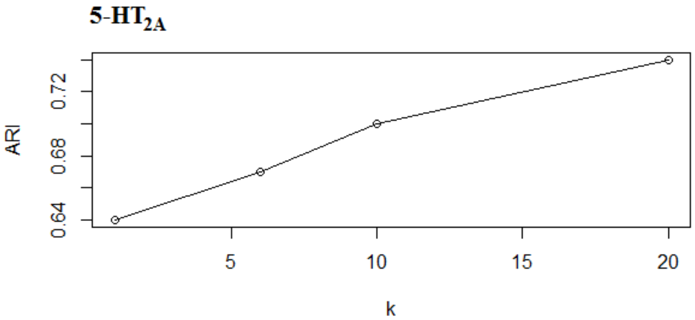
k-means

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

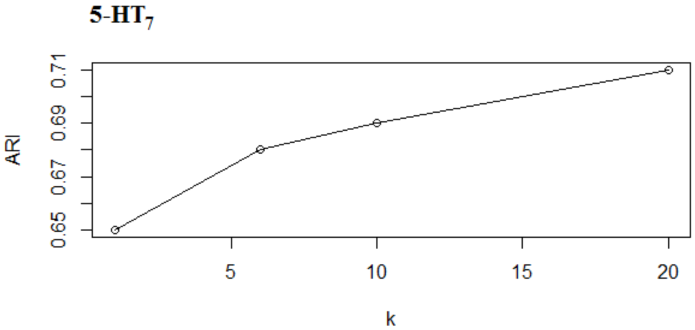
hierarchical clustering

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

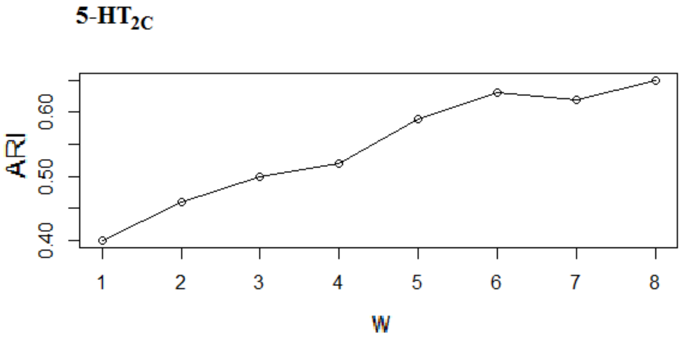
after optimization process



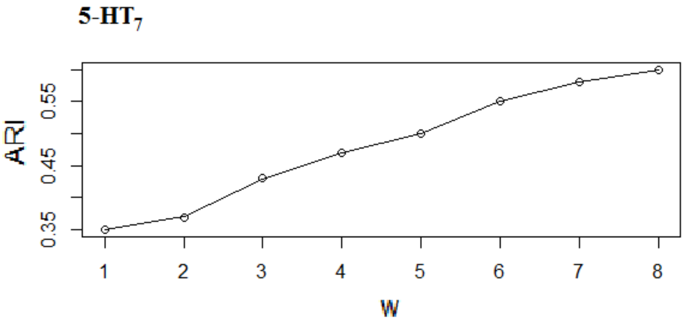
after optimization process



more explanatory variables



more explanatory variables



Conclusion

- metric learning problem
- a single function which combines data representation-metric pairs can improve the performance of metric-based algorithms

References

Aczel A., Sounderpandian J.; *Complete Business Statistics, McGraw Hill, NY, 2009.*

Atkeson C., Moore A., Schaal S; Locally weighted learning, Artificial Intelligence Review 11, 1997, pp. 11-73.

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

Roweis S.T., Saul L.K.; Nonlinear dimensionality reduction by locally linear embedding, Science 290, 2000, pp. 2323-2326.

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.

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