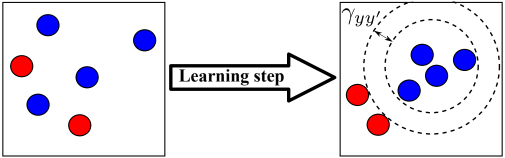
A Theoretical Analysis of Metric Hypothesis Transfer Learning

Michaël Perrot[1] and Amaury Habrard[1]
{michael.perrot,amaury.habrard}@univ-st-etienne.fr

[1]Université de Lyon, Université Jean Monnet de Saint-Etienne,
Laboratoire Hubert Curien, CNRS, UMR5516, F-42000, Saint-Etienne, France.

Metric Learning

Learning how to compare objects : learn a new space where some constraints are fulfilled, e.g. move closer circles of the same color (class) and keep far away circles of different colors (classes).



Mahalanobis-like Distance

$$d_{\mathbf{M}}(\mathbf{x},\mathbf{x}') = \sqrt{(\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}')},\ \mathbf{M} \text{ a PSD matrix } (\mathbf{M} = \mathbf{L}\mathbf{L}^T).$$

Well-known distances

- Euclidean Distance : $\mathbf{M} = \mathbf{I}$
- Original Mahalanobis Distance : $\mathbf{M} = \mathbf{\Sigma}^{-1}$
- Zero Distance : $\mathbf{M} = \mathbf{0}$

Regularized Metric Learning

$$\arg\min_{\mathbf{M}\succeq 0} L_T(\mathbf{M}) + \lambda\|\mathbf{M}\|_{\mathcal{F}}^2 \quad (1)$$

with :

- $T = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset (\mathcal{X} \times \mathcal{Y})^n$, a learning sample
- $L_T(\mathbf{M}) = \frac{1}{n^2}\sum_{\mathbf{z},\mathbf{z}'\in T} l(\mathbf{M},\mathbf{z},\mathbf{z}')$
  with $l(\mathbf{M},\mathbf{z},\mathbf{z}')$:
  - convex with respect to $\mathbf{M}$
  - (σ, m)-admissible
  - k-lipschitz
  - penalizing high distances between similar examples et small distances between dissimilar examples
- $\|\cdot\|_{\mathcal{F}}$, the Frobenius norm

Regularized Metric Learning

$$\arg\min_{\mathbf{M}\succeq 0} L_T(\mathbf{M}) + \lambda\|\mathbf{M} - \mathbf{0}\|_{\mathcal{F}}^2 \qquad (1)$$

with :

- $T = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset (\mathcal{X} \times \mathcal{Y})^n$, a learning sample
- $L_T(\mathbf{M}) = \frac{1}{n^2}\sum_{\mathbf{z},\mathbf{z}'\in T} l(\mathbf{M},\mathbf{z},\mathbf{z}')$
  with $l(\mathbf{M},\mathbf{z},\mathbf{z}')$:
  - convex with respect to **M**
  - (σ, m)-admissible
  - k-lipschitz
  - penalizing high distances between similar examples et small distances between dissimilar examples
- $\|\cdot\|_{\mathcal{F}}$, the Frobenius norm

Biased Regularized Metric Learning

$$\arg\min_{\mathbf{M}\succeq 0} L_T(\mathbf{M}) + \lambda \|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \quad (1)$$

with :

- $T = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset (\mathcal{X} \times \mathcal{Y})^n$, a learning sample
- $L_T(\mathbf{M}) = \frac{1}{n^2} \sum_{\mathbf{z},\mathbf{z}' \in T} l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$
  with $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$:
  - convex with respect to $\mathbf{M}$
  - (σ, m)-admissible
  - k-lipschitz
  - penalizing high distances between similar examples et small distances between dissimilar examples
- $\| \cdot \|_{\mathcal{F}}$, the Frobenius norm
- $\mathbf{M}_{\mathcal{S}}$, a fixed metric biasing the regularization,
  e.g. $\mathbf{I}, \mathbf{\Sigma}^{-1}$, a metric learned from another domain, ...

Biased Regularized Metric Learning

$$\arg\min_{\mathbf{M}\succeq 0} L_T(\mathbf{M}) + \lambda\|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \quad (1)$$

with :

- $T = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset (\mathcal{X} \times \mathcal{Y})^n$, a learning sample
- $L_T(\mathbf{M}) = \frac{1}{n^2}\sum_{\mathbf{z},\mathbf{z}' \in T} l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$
  with $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$:
  - convex with respect to **M**
  - (σ, m)-admissible
  - k-lipschitz
  - penalizing high distances between similar examples et small distances between dissimilar examples
- $\|\cdot\|_{\mathcal{F}}$, the Frobenius norm
- $\mathbf{M}_{\mathcal{S}}$, a fixed metric biasing the regularization,
  e.g. $\mathbf{I}, \mathbf{\Sigma}^{-1}$, a metric learned from another domain, ...

Hypothesis Transfer Learning has already been studied in a different setting [Kuzborskij and Orabona, 2013, 2014].

Biased Regularized Metric Learning

$$\arg\min_{\mathbf{M}\succeq 0} L_T(\mathbf{M}) + \lambda\|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \tag{1}$$

with :

- $T = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset (\mathcal{X} \times \mathcal{Y})^n$, a learning sample
- $L_T(\mathbf{M}) = \frac{1}{n^2}\sum_{\mathbf{z},\mathbf{z}' \in T} l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$
  with $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$:
  - convex with respect to **M**
  - (σ, m)-admissible
  - k-lipschitz
  - penalizing high distances between similar examples et small distances between dissimilar examples
- $\|\cdot\|_{\mathcal{F}}$, the Frobenius norm
- $\mathbf{M}_{\mathcal{S}}$, a fixed metric biasing the regularization,
  e.g. $\mathbf{I}, \mathbf{\Sigma}^{-1}$, a metric learned from another domain, ...

Objective : Provide a theoretical analysis of biased regularized metric learning and propose an efficient way to reweight the source metric.

General Definitions

(σ, m)-admissibility

A loss function is (σ, m)-admissible for metric learning if the loss difference between two pairs of examples is bounded by a constant σ times a quantity only related to the labels plus a constant :

$$|l(\mathbf{M}, \mathbf{z}_1, \mathbf{z}_2) - l(\mathbf{M}, \mathbf{z}_3, \mathbf{z}_4)| \leq \sigma\, |y_1 y_2 - y_3 y_4| + m.$$

k-lipschitz continuity

A loss function is k-lipschitz continuous if the loss difference between two metrics is bounded by a constant k times a quantity which only depends on the difference between the two metrics :

$$\left| l(\mathbf{M}, \mathbf{z}, \mathbf{z}') - l(\mathbf{M}', \mathbf{z}, \mathbf{z}') \right| \leq k \|\mathbf{M} - \mathbf{M}'\|_{\mathcal{F}}.$$

On Average Replace Two Stability

The expected loss difference when replacing two examples in the training set is bounded by a value decreasing in $\mathcal{O}\left(\frac{1}{n}\right)$.

Extension to metric learning of [Shalev-Shwartz et al., 2010].

Definition (On-average-replace-two-stability)

Let $\epsilon : \mathbb{N} \rightarrow \mathbb{R}$ be monotonically decreasing and let $U(n)$ be the uniform distribution over $\{1 \dots n\}$. A metric learning algorithm is on-average-replace-two-stable with rate $\epsilon(n)$ if for every distribution $\mathcal{D}_\mathcal{T}$:

$$\mathbb{E}_{\substack{T \sim \mathcal{D}_\mathcal{T}{}^n \\ i,j \sim U(n) \\ \mathbf{z}_1, \mathbf{z}_2 \sim \mathcal{D}_\mathcal{T}}} \left[l(\mathbf{M}^{ij^*}, \mathbf{z}^i, \mathbf{z}^j) - l(\mathbf{M}^*, \mathbf{z}^i, \mathbf{z}^j) \right] \leq \epsilon(n)$$

where $\mathbf{M}^*$, respectively $\mathbf{M}^{ij^*}$, is the optimal solution when learning with the training set T, respectively T^{ij}. T^{ij} is obtained by replacing $\mathbf{z}^i$, the i^{th} example of T, by $\mathbf{z}_1$ to get a training set T^i and then by replacing $\mathbf{z}^j$, the j^{th} example of T^i, by $\mathbf{z}_2$.

On Average Bound

The learned metric is on average at least as good as the source metric.

Theorem (On-average-replace-two-stability)

Given a training sample T of size n drawn i.i.d. from $\mathcal{D}_\mathcal{T}$, an algorithm solving optimization problem (1) *is on-average-replace-two-stable with* $\epsilon(n) = \frac{8k^2}{\lambda n}$.

On Average Bound

The learned metric is on average at least as good as the source metric.

Theorem (On-average-replace-two-stability)

Given a training sample T of size n drawn i.i.d. from $\mathcal{D}_{\mathcal{T}}$, an algorithm solving optimization problem (1) *is on-average-replace-two-stable with* $\epsilon(n) = \frac{8k^2}{\lambda n}$.

Theorem (On average bound)

For any convex, k-lipschitz loss, we have :

$$\mathbb{E}_{T \sim \mathcal{D}_{\mathcal{T}}{}^n} [L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}^*)] \leq L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}_{\mathcal{S}}) + \frac{8k^2}{\lambda n}$$

where the expected value is taken over size-n training sets.

Uniform Stability

Changing an example in the training set does not change much the outcome of the algorithm.

Definition (Uniform stability [Bousquet and Elisseeff, 2002, Jin et al., 2009])

An algorithm has a uniform stability in $\epsilon(n)$ *if* $\forall i$,

$$\sup_{\mathbf{z},\mathbf{z}' \sim \mathcal{D}_{\mathcal{T}}} \left| l(\mathbf{M}^*, \mathbf{z}, \mathbf{z}') - l(\mathbf{M}^{i^*}, \mathbf{z}, \mathbf{z}') \right| \leq \epsilon(n)$$

where $\mathbf{M}^*$ *is the matrix learned on the training set* T *and* $\mathbf{M}^{i^*}$ *is the matrix learned on the training set* T^i *obtained by replacing the* i^{th} *example of* T *by a new independent one.*

Generalisation Bound

The biased regularized metric learning framework is consistent.

Theorem (Uniform stability)

Given a training sample T of n examples drawn i.i.d. from $\mathcal{D}_{\mathcal{T}}$, an algorithm solving optimization problem (1) *has a uniform stability in* $\epsilon(n) = \frac{4k^2}{\lambda n}$.

Generalisation Bound

The biased regularized metric learning framework is consistent.

Theorem (Uniform stability)

Given a training sample T of n examples drawn i.i.d. from $\mathcal{D}_\mathcal{T}$, an algorithm solving optimization problem (1) *has a uniform stability in* $\epsilon(n) = \frac{4k^2}{\lambda n}$.

Theorem (Generalization bound)

With probability $1 - \delta$, *for any matrix* $\mathbf{M}^*$ *learned with an* $\epsilon(n)$ *uniformly stable algorithm and for any convex, k-lipschitz and* (σ, m)*-admissible loss, we have :*

$$L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) \leq L_T(\mathbf{M}^*) + (4\sigma + 2m + c)\sqrt{\frac{\ln \frac{2}{\delta}}{2n}} + \mathcal{O}\left(\frac{1}{n}\right)$$

where c *is a constant linked to the k-lipschitz property of the loss and* $\epsilon(n)$ *appears in* $\mathcal{O}\left(\frac{1}{n}\right)$.

Application to a Specific Loss

We consider the following loss (inspired from [Jin et al., 2009]) :

$$l(\mathbf{M},\mathbf{z},\mathbf{z}') = \left[yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') - \gamma_{yy'})\right]_+ \qquad (2)$$

where $[\cdot]_+$ is the hinge loss, $yy' = 1$ for examples of the same class and -1 otherwise and $\gamma_{yy'}$ is the chosen margin.

Lemma ((σ, m)-admissibility)

Let $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$ be four examples and $\mathbf{M}^*$ be the optimal solution of Problem 1. The convex and k-lipschitz loss function $l(\mathbf{M},\mathbf{z},\mathbf{z}')$ is (σ, m)-admissible with $\sigma = \max(\gamma_{y_3y_4}, \gamma_{y_1y_2})$ and $m = 2\max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x}-\mathbf{x}'\|^2 \left(\sqrt{\frac{L_T(\mathbf{M}_\mathcal{S})}{\lambda}} + \|\mathbf{M}_\mathcal{S}\|_\mathcal{F}\right)$.

Application to a Specific Loss

We consider the following loss (inspired from [Jin et al., 2009]) :

$$l(\mathbf{M}, \mathbf{z}, \mathbf{z}') = \left[yy'((\mathbf{x} - \mathbf{x}')^T \mathbf{M}(\mathbf{x} - \mathbf{x}') - \gamma_{yy'}) \right]_+ \quad (2)$$

where $[\cdot]_+$ is the hinge loss, $yy' = 1$ for examples of the same class and -1 otherwise and $\gamma_{yy'}$ is the chosen margin.

Theorem (Generalization bound)

With probability $1 - \delta$ *for any matrix* $\mathbf{M}^*$ *learned by an algorithm solving optimization problem* (1) *with loss* (2)*, we have :*

$$L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) \leq L_\mathcal{T}(\mathbf{M}^*) + 4 \left(\sqrt{\frac{L_\mathcal{T}(\mathbf{M}_\mathcal{S})}{\lambda}} + \|\mathbf{M}_\mathcal{S}\|_\mathcal{F} + c_\gamma \right) \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} + \mathcal{O}\left(\frac{1}{n}\right)$$

where c_γ *is a constant linked to the k-lipschitz property of the loss and the chosen margins.*

Reweighting the Source Metric

Let $\mathbf{M}_{\mathcal{S}} = \beta \mathbf{M}_{\mathbf{SOURCE}}$, we want to minimize the right hand side of the bound, i.e. to choose the best matrix to transfer. Hence, we search β such that :

$$\beta^* = \arg\min_{\beta} \sqrt{\frac{L_T(\beta \mathbf{M}_{\mathbf{SOURCE}})}{\lambda}} + \|\beta \mathbf{M}_{\mathbf{SOURCE}}\|_{\mathcal{F}} \quad (3)$$

Reweighting the Source Metric

Let $\mathbf{M}_{\mathcal{S}} = \beta \mathbf{M}_{\mathbf{SOURCE}}$, we want to minimize the right hand side of the bound, i.e. to choose the best matrix to transfer. Hence, we search β such that :

$$\beta^* = \arg\min_{\beta} \sqrt{\frac{L_T(\beta \mathbf{M}_{\mathbf{SOURCE}})}{\lambda}} + \|\beta \mathbf{M}_{\mathbf{SOURCE}}\|_{\mathcal{F}} \tag{3}$$

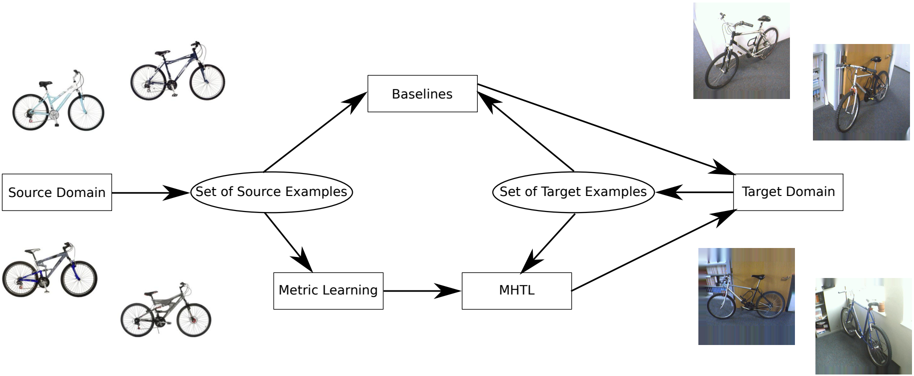
Interest of Tuning β

	Baselines		Solving optimization problem (1) with loss (2)			
Dataset	1-NN	ITML	$\mathbf{M}_{\mathcal{S}} = \beta \mathbf{I}$	$\mathbf{M}_{\mathcal{S}} = \mathbf{I}$	$\mathbf{M}_{\mathcal{S}} = \beta \mathbf{\Sigma}^{-1}$	$\mathbf{M}_{\mathcal{S}} = \mathbf{\Sigma}^{-1}$
Breast	95.31 ± 1.11	95.40 ± 1.37	**96.06 ± 0.77**	95.75 ± 0.87	95.71 ± 0.84	94.76 ± 1.38
Pima	67.92 ± 1.95	68.13 ± 1.86	67.87 ± 1.57	67.54 ± 1.99	**68.37 ± 2.00**	66.31 ± 2.37
Scale	78.73 ± 1.69	**87.31 ± 2.35**	80.98 ± 1.51	80.82 ± 1.27	81.35 ± 1.17	80.88 ± 1.43
Wine	93.40 ± 2.70	93.82 ± 2.63	**95.42 ± 1.71**	95.07 ± 1.68	94.31 ± 2.01	80.56 ± 5.75

Application to a Transfer Learning Task

Setting

The idea is to learn a metric on a source domain and to use this metric to bias the regularizer when learning on the target domain.



MHTL : Metric Hypothesis Transfer Learning

Application to a Transfer Learning Task

Setting

The idea is to learn a metric on a source domain and to use this metric to bias the regularizer when learning on the target domain.

On the Office-Caltech dataset

	Baselines			Solving optimization problem (1) with loss (2)		
Task	1-NN$_{\mathcal{S}}$	MMDT	GFK	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{\Sigma}^{-1}$	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{M}_{\mathbf{ITML}}$	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{M}_{\mathbf{LMNN}}$
A $\rightarrow$ C	35.95 $\pm$ 1.30	**39.76 $\pm$ 2.25**	37.81 $\pm$ 1.85	32.65 $\pm$ 3.76	32.93 $\pm$ 4.60	34.66 $\pm$ 3.66
A $\rightarrow$ D	33.58 $\pm$ 4.37	54.25 $\pm$ 4.32	51.54 $\pm$ 3.55	54.69 $\pm$ 3.96	51.54 $\pm$ 4.03	**54.72 $\pm$ 5.00**
A $\rightarrow$ W	33.68 $\pm$ 3.60	64.91 $\pm$ 5.71	59.36 $\pm$ 4.30	67.11 $\pm$ 5.11	64.09 $\pm$ 5.20	**67.62 $\pm$ 5.18**
C $\rightarrow$ A	37.37 $\pm$ 2.95	**51.05 $\pm$ 3.38**	46.36 $\pm$ 2.94	50.15 $\pm$ 4.87	49.89 $\pm$ 5.25	50.36 $\pm$ 4.67
C $\rightarrow$ D	31.89 $\pm$ 5.77	52.80 $\pm$ 4.84	**58.07 $\pm$ 3.90**	56.77 $\pm$ 4.63	53.78 $\pm$ 7.23	57.44 $\pm$ 4.48
C $\rightarrow$ W	28.60 $\pm$ 6.13	62.75 $\pm$ 5.19	63.26 $\pm$ 5.89	64.64 $\pm$ 6.44	64.00 $\pm$ 6.08	**65.11 $\pm$ 5.25**
D $\rightarrow$ A	33.59 $\pm$ 1.77	**50.39 $\pm$ 3.40**	40.77 $\pm$ 2.55	49.48 $\pm$ 4.41	49.11 $\pm$ 4.09	49.67 $\pm$ 4.00
D $\rightarrow$ C	31.16 $\pm$ 1.19	**35.70 $\pm$ 3.25**	30.64 $\pm$ 1.98	32.90 $\pm$ 3.14	32.99 $\pm$ 3.58	33.84 $\pm$ 2.99
D $\rightarrow$ W	**76.92 $\pm$ 2.18**	74.43 $\pm$ 3.10	74.98 $\pm$ 2.89	65.57 $\pm$ 4.52	66.38 $\pm$ 6.04	69.72 $\pm$ 3.78
W $\rightarrow$ A	32.19 $\pm$ 3.04	50.56 $\pm$ 3.66	43.26 $\pm$ 2.34	50.80 $\pm$ 3.63	50.16 $\pm$ 4.32	**50.92 $\pm$ 4.00**
W $\rightarrow$ C	27.67 $\pm$ 2.58	**34.86 $\pm$ 3.62**	29.95 $\pm$ 3.05	31.54 $\pm$ 3.60	31.40 $\pm$ 4.29	32.64 $\pm$ 3.52
W $\rightarrow$ D	64.61 $\pm$ 4.30	62.52 $\pm$ 4.40	**71.93 $\pm$ 4.07**	57.17 $\pm$ 6.50	56.85 $\pm$ 5.51	61.14 $\pm$ 5.78
Mean	38.93 $\pm$ 3.26	**52.83 $\pm$ 3.93**	50.66 $\pm$ 3.28	51.12 $\pm$ 4.55	50.26 $\pm$ 5.02	52.32 $\pm$ 4.36

MHTL, using only the source metric, is competitive with the baselines.

Conclusion and Perspectives

We proposed a study of Biased Regularized Metric Learning through :

- An On Average analysis showing that with a fast convergence rate the learned metric is better than the source metric.
- A Consistency Analysis proving that biasing the regularization term toward a source metric does not challenge the consistency of the approach.
- A Reweighting Algorithm allowing us to weight the source metric with respect to the problem at hand when we consider a specific loss.

Conclusion and Perspectives

We proposed a study of Biased Regularized Metric Learning through :

- An On Average analysis showing that with a fast convergence rate the learned metric is better than the source metric.
- A Consistency Analysis proving that biasing the regularization term toward a source metric does not challenge the consistency of the approach.
- A Reweighting Algorithm allowing us to weight the source metric with respect to the problem at hand when we consider a specific loss.

A perspective of this work would be to extend the framework to other settings and other kind of regularizers.

References I

Bellet, A., Habrard, A., and Sebban, M. (2015).
Metric Learning.
Morgan & Claypool Publishers.

Bousquet, O. and Elisseeff, A. (2002).
Stability and generalization.
Journal of Machine Learning Research, 2 :499–526.

Jin, R., Wang, S., and Zhou, Y. (2009).
Regularized distance metric learning : Theory and algorithm.
In *Proc. of NIPS*, pages 862–870.

Kuzborskij, I. and Orabona, F. (2013).
Stability and hypothesis transfer learning.
In *Proc. of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 942–950.

References II

Kuzborskij, I. and Orabona, F. (2014).
Learning by transferring from auxiliary hypotheses.
CoRR, abs/1412.1619.

Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding Machine Learning : From Theory to Algorithms.
Cambridge University Press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010).
Learnability, stability and uniform convergence.
Journal of Machine Learning Research, 11 :2635–2670.