Similarity-based Learning via Data Driven Embeddings

Purushottam Kar[1] and Prateek Jain[2]

[1]Indian Institute of Technology, Kanpur, UP, INDIA
[2]Microsoft Research Lab, Bangalore, KA, INDIA

Microsoft Research

Abstract

- Proliferation of machine learning algorithms in diverse domains
  - necessitates working with non-explicit features
  - notions of distance/similarity more natural than hand-coded features Co-authorship graphs, Earth-mover's distance
  - Typically end up with non-PSD similarity measures (kernels)
- **Goal** : a model of learning with arbitrary similarity measures
- **Our Contributions** :
  - develop a general notion of *goodness* for similarity measures
  - propose algorithms that make optimal* use of any such measure
  - provide classifiers with provable error bounds

Existing Work

- Models of learning using similarity/distance functions [1, 2]
- Direct use of indefinite kernels with SVMs [3]
- Use of similarities as features as against kernels [4]
- Can we take into account the suitability of these measures ?
  - [1] defines a notion of suitability for similarities (BBS)
  - [2] gives similar treatement to distances (DBOOST)
  - Yield classifiers with bounded generalization error
- **This paper** : a model with more flexible notion of suitability for similarities
- All our results hold for (non-metric) distance functions as well

What is a *good* similarity function ?

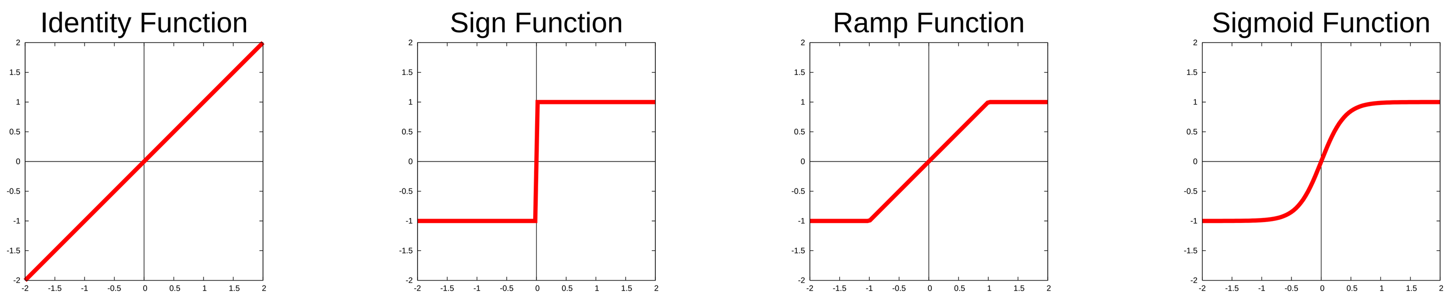
- Suitability of a similarity function to a given classification problem
  - Points with same label should be more similar than dissimilarly labeled points
  - Link a formal notion of suitability to error bounds (agnostic learning) [1]
  - Generalize the notion of suitability to be data dependent for more flexibility

Definition 1. (Good Similarity Function)

A similarity function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is (ϵ, γ, B)-good for a classification problem if for some antisymmetric transfer function $f : \mathbb{R} \rightarrow [-1, 1]$ and a weight function $w : \mathcal{X} \times \mathcal{X} \rightarrow [-B, B]$, at least a $(1 - \epsilon)$ fraction of examples $x \sim \mathcal{D}$ satisfies $G_f(x) \geq \gamma$ where

$$G_f(x) = \mathop{\mathbb{E}}_{\substack{x' \sim \mathcal{D}, \ell(x') = \ell(x) \\ x'' \sim \mathcal{D}, \ell(x'') \neq \ell(x)}} \left[w(x', x'') f(K(x, x') - K(x, x'')) \right].$$

Examples of Transfer functions



- This model encompasses BBS and DBOOST
  - BBS uses identity, DBOOST uses sign as transfer function

References

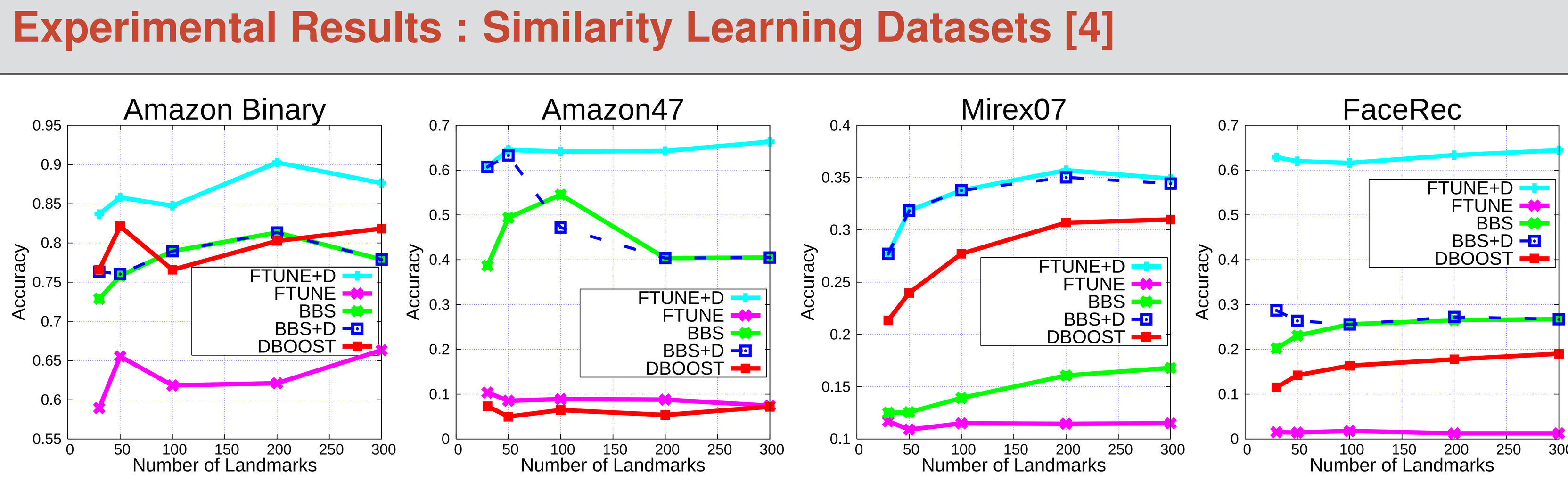
[1] Maria-Florina Balcan and Avrim Blum. On a Theory of Learning with Similarity Functions. In *International Conference on Machine Learning*, pages 73–80, 2006.

[2] Liwei Wang, Cheng Yang, and Jufu Feng. On Learning with Dissimilarity Functions. In *International Conference on Machine Learning*, pages 991–998, 2007.

[3] Bernard Haasdonk. Feature Space Interpretation of SVMs with Indefinite Kernels. *IEEE Transactions on Pattern Analysis and Machince Intelligence*, 27(4):482–492, 2005.

[4] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based Classification: Concepts and Algorithms. *J. Machine Learning Research*, 10:747–776, 2009.

Experimental Results : Similarity Learning Datasets [4]



Dataset	BBS	DBOOST	FTUNE+D-S
Amazon-Bin	0.73	0.77	0.84
AuralSonar	0.82	0.81	0.80
Patrol	0.51	0.34	0.58
Voting	0.95	0.94	0.94
Protein	0.98	1.00	0.98
Mirex07	0.12	0.21	0.28
Amazon47	0.39	0.07	0.61
FaceRec	0.20	0.12	0.63

Dataset	BBS	DBOOST	FTUNE+D-S
Amazon-Bin	0.78	0.82	0.88
AuralSonar	0.88	0.85	0.85
Patrol	0.79	0.55	0.79
Voting	0.97	0.97	0.97
Protein	0.98	0.99	0.98
Mirex07	0.17	0.31	0.35
Amazon47	0.40	0.07	0.66
FaceRec	0.27	0.19	0.64

- Using validation to choose f leads to overfitting (average dataset size 660)
- DSELECT removes redundancies and chooses informative set of landmarks

Learning Algorithm

- **Given** : A kernel K, a transfer function f and landmark pairs $\mathcal{L} = (x_i^+, x_i^-)_{i=1}^n$
  - All x_i^+ are positively labeled and all x_i^- are negatively labeled
- For any $x \in \mathcal{X}$, define $\Phi_{\mathcal{L}}(x) = \begin{bmatrix} f(K(x, x_1^+) - K(x, x_1^-)) \\ f(K(x, x_2^+) - K(x, x_2^-)) \\ \cdot \\ \cdot \\ \cdot \\ f(K(x, x_n^+) - K(x, x_n^-)) \end{bmatrix} \in \mathbb{R}^n$
- Learn a hyperplane in $\mathbb{R}^n$ using a training set T
  - $\ell_{\text{lin}} \leftarrow$ LEARN-LINEAR($\mathcal{L}(T)$)
  - LEARN-LINEAR may be taken to be L1-SVM, LR, Perceptron ...
  - guarantees allow use of any Lipschitz loss function - hinge, logit, quadratic ...
- Output $\hat{\ell} : \mathcal{X} \rightarrow \{-1, +1\}$ defined as $\hat{\ell} : x \mapsto \ell_{\text{lin}}(\Phi_{\mathcal{L}}(x))$

Generalization Guarantee

- Modify Definition 1 to include a loss function L : require $L(f) := \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [L(G_f(x))] \leq \epsilon$
  - Definition 1 can be shown to use the loss function $L(x) = \mathbb{1}_{\{x < \gamma\}}$

Theorem 2. (Generalization Guarantee)

If K is an (ϵ, B)-good similarity function with respect to a C-Lipschitz loss function L then for any $\epsilon_1, \delta > 0$, taking $n = \frac{16B^2C^2}{\epsilon_1^2} \ln\left(\frac{4B}{\delta\epsilon_1}\right)$ random landmark pairs suffice to output a classifier with expected L-loss less than $\epsilon + \epsilon_1$ with probability $1 - \delta$.

- Guarantees the existence of a good linear classifier in $\mathbb{R}^n$ if f is *suitable*
- Missing pieces
  - How to find a good f from a given family $\mathcal{F}$?
  - Better than random choice of landmark pairs ?

Selecting a good transfer function

- Goodness of a transfer function f quantified using $L(f)$
- Let $L(f, \mathcal{L})$ be the L-loss of the best classifier that uses the landmarks set $\mathcal{L}$
- Theorem 2 guarantees $L(f, \mathcal{L}) \leq L(f) + \epsilon_1$ for a fixed transfer function f

Theorem 3. (Uniform Convergence Bound)

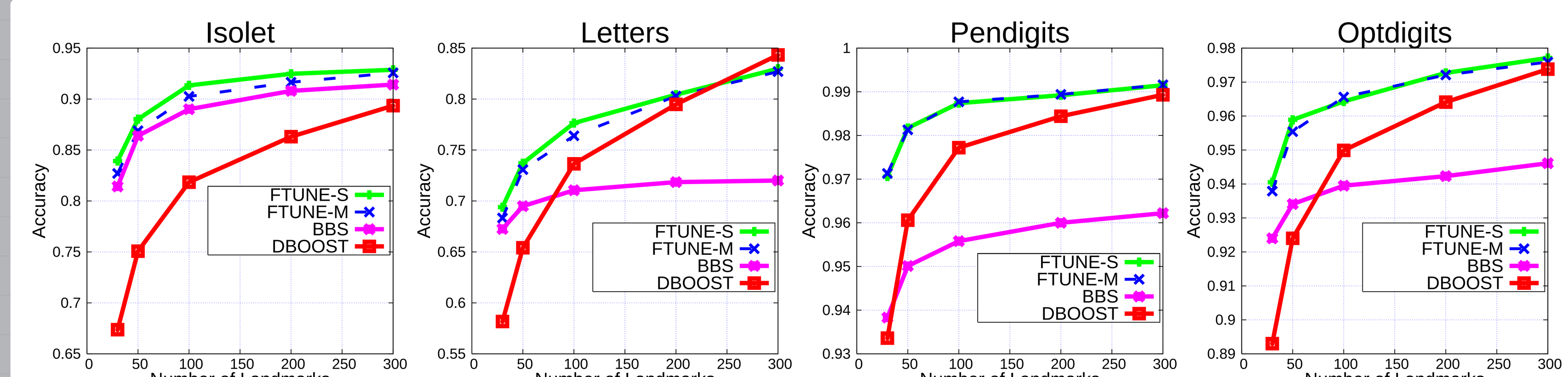
If $\mathcal{F}$ is a set of transfer functions with an ϵ-net with respect to infinity norm at scale $r = \frac{\epsilon_1}{4C_L B}$ of size atmost $\mathcal{N}(\mathcal{F}, r)$, then for any $\epsilon_1, \delta > 0$, $n = \frac{64B^2C_L^2}{\epsilon_1^2} \ln\left(\frac{16B \cdot \mathcal{N}(\mathcal{F}, r)}{\delta\epsilon_1}\right)$ random landmark pairs ensure $\sup_{f \in \mathcal{F}} [|L(f, \mathcal{L}) - L(f)|] \leq \epsilon_1$ with probability $1 - \delta$.

- Guarantees that suitability of f will be evident in $\mathbb{R}^n$ for *all* $f \in \mathcal{F}$ with a single $\mathcal{L}$
- Validates the use of ERM style algorithms to select a good f from $\mathcal{F}$
- e.g. possible to tune paramter σ in sigmoid transfer function

Landmark Selection

- On small datasets, choice of transfer function can lead to overfitting
- DSELECT: heuristic for landmark selection that improves performance
  - If landmarks clumped together then all training points get same embedding
  - Need to promote *diversity* among landmark points
- Incrementally select landmark points in a greedy manner
  - At each step choose a point that is *least similar* to already chosen points
  - Form pairs out of these points later on to get landmark pairs

Experimental Results : UCI Benchmark Datasets



Dataset	BBS	DBOOST	FTUNE-S	FTUNE-M
Cod-rna	0.93	0.89	0.93	0.93
Isolet	0.81	0.67	0.84	0.83
Letters	0.67	0.58	0.69	0.68
Magic	0.82	0.81	0.84	0.84
Pen-digits	0.94	0.93	0.97	0.97
Nursery	0.91	0.91	0.90	0.90
Faults	0.70	0.68	0.70	0.71
Mfeat-px	0.94	0.91	0.95	0.94
Mfeat-zn	0.79	0.72	0.79	0.79
Opt-digits	0.92	0.89	0.94	0.94
Satellite	0.85	0.86	0.86	0.87
Segment	0.90	0.93	0.92	0.92

Dataset	BBS	DBOOST	FTUNE-S	FTUNE-M
Cod-rna	0.94	0.93	0.94	0.94
Isolet	0.91	0.89	0.93	0.93
Letters	0.72	0.84	0.83	0.83
Magic	0.84	0.84	0.85	0.85
Pen-digits	0.96	0.99	0.99	0.99
Nursery	0.93	0.97	0.96	0.97
Faults	0.72	0.74	0.73	0.73
Mfeat-px	0.96	0.97	0.97	0.97
Mfeat-zn	0.81	0.79	0.82	0.82
Opt-digits	0.95	0.97	0.98	0.98
Satellite	0.85	0.90	0.89	0.89
Segment	0.90	0.96	0.96	0.96

- Average dataset size 13200 : validation can be performed without overfitting
- DSELECT does not help on large datasets : FTUNE alone performs well