Learning Better Data Representation using Inference Driven Metric Learning (IDML)

Paramveer S. Dhillon[1], Partha Pratim Talukdar[2] and Koby Crammer[3]

[1]CIS, University of Pennsylvania, Philadelphia, PA, U.S.A. [2]Search Labs, Microsoft Research, Mountain View, CA, U.S.A. [3]Electrical Engineering, The Technion, Haifa, Israel

High Dimensional Data in NLP

Problem:
- NLP datasets are high dimensional.
- Reliable parameter estimation from high dimensional data is hard.
- Better data representation is needed for effective learning!

Existing Solution:
1. Project data into a lower dimensional space using *unsupervised dimensionality reduction* methods (e.g. PCA, NMF).
2. Perform learning in the lower dimensional space.

- However, a limited amount of labeled data, along with vast amounts of unlabeled data are also available.
- Recently proposed metric learning algorithms [1, 2] make use of such resources to learn a (Mahalanobis) distance metric.
- These metric learning algorithms learn a representation of the data (more below).

We explore the effectiveness of data representations learned by metric learning algorithms for NLP tasks.

Metric Learning & Linear Projection

Distance metric learning algorithms [1, 2] *learn* the Mahalanobis distance, $d_A(x_i, x_j)$, between instances x_i and x_j.

$$d_A(x_i, x_j) = (x_i - x_j)^\top A (x_i - x_j)$$

where A is s $d \times d$ positive-definite matrix, and d is the data dimension. This matrix can be written as $A = P^\top P$, where P is another matrix of size $d \times d$.

$$\begin{aligned} d_A(x_i, x_j) &= (x_i - x_j)^\top A (x_i - x_j) \\ &= (x_i - x_j)^\top P^\top P (x_i - x_j) \\ &= (Px_i - Px_j)^\top (Px_i - Px_j) \\ &= d_{\text{Euclidean}}(Px_i, Px_j) \end{aligned}$$

- Hence, computing Mahalanobis distance in the original space is equivalent to computing Euclidean distance in the space induced by projection matrix P.
- **Learning can now be performed in the data representation induced by P.**

Distance Metric Learning Algorithms

- Supervised distance metric learning
  - Uses labeled data to learn a Mahalanobis metric.
  - Goal is to reduce distance between instances with the same label, while separating instances with different labels.
  - Information Theoretic Metric Learning (ITML) [1].
  - Large Margin Nearest Neighbors (LMNN) [3].

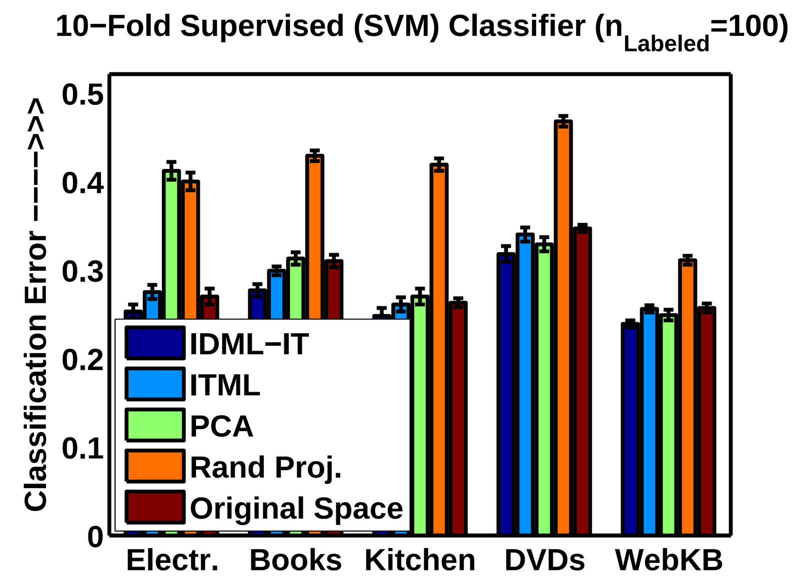
- Semi-Supervised distance metric learning
  - Learns Mahalanobis metric using labeled as well as *unlabeled* data.
  - Inference Driven Metric Learning (IDML) [2].
    - Incorporates unlabeled data via self-training.
    - Adds high-confidence (low entropy) instances with their predicted labels for the next iteration of metric learning.
    - Can be used in conjunction with any supervised metric learning algorithm (like LMNN, ITML).
    - Enables to learn a better metric as the amount of supervision is increased.
    - Guaranteed to converge, as number of available unlabeled instances decrease at each iteration, or else there are no more instances to add.

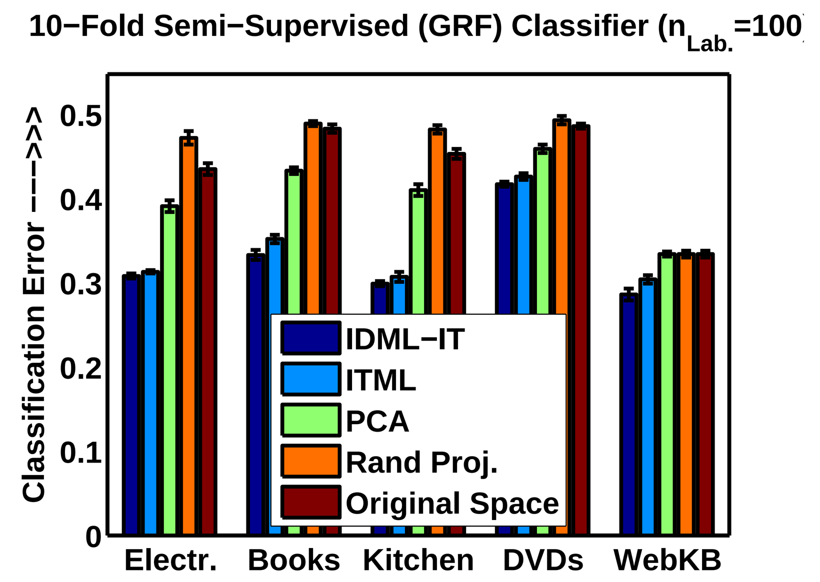
Algorithm 1 Inference Driven Metric Learning (IDML)
Input: instances X, training labels Y, training instance indicator S, label entropy threshold β, neighborhood size k
Output: Mahalanobis distance parameter A

$\hat{Y} \leftarrow Y$, $\hat{S} \leftarrow S$
repeat
  $A \leftarrow \text{METRICLEARNER}(X, \hat{S}, \hat{Y})$
  $W \leftarrow \text{CONSTRUCTKNNGRAPH}(X, A, k)$
  $\hat{Y}' \leftarrow \text{GRAPHLABELINFERENCE}(W, \hat{S}, \hat{Y})$
  $U \leftarrow \text{SELECTLOWENTINSTANCES}(\hat{Y}', \hat{S}, \beta)$
  $\hat{Y} \leftarrow \hat{Y} + U\hat{Y}'$
  $\hat{S} \leftarrow \hat{S} + U$
until convergence (i.e. $U_{ii} = 0, \ \forall i$)
return A

Experimental Results

- Classification accuracies on 5 Text Classification and Sentiment Analysis Datasets, with 1500 total instances for all datasets, and 100 labeled instances. (more results in the paper)
- Data representations compared:
  - Original Space: No data transformation
  - PCA Space: PCA applied to the original space, resulting in a 250 dimensional space
  - Random Projections: Random projection applied to the original space
  - ITML: Projection matrix P learned by ITML [1]
  - IDML-IT: Projection matrix P learned by IDML with ITML as the METRICLEARNER (see Algorithm 2), in the transductive setting.
- Classifiers used:
  - SVM with RBF kernel, and Gaussian Random Field (GRF) [4]





Summary

Distance metric learning helps in learning better data representations, resulting in better classification accuracies.

IDML is likely to be useful for high dimensional data, e.g., problems in NLP, as demonstrated through the experiments on Text Classification and Sentiment Analysis datasets.

References

[1] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.
[2] P. S. Dhillon, P. P. Talukdar, and K. Crammer. Inference-driven metric learning for graph construction. Technical Report MS-CIS-10-18, CIS Department, University of Pennsylvania, May 2010.
[3] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 2009.
[4] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.