An Overview of Distance Metric Learning

Liu Yang

October 28, 2007

In our previous comprehensive survey [41], we have categorized the disparate issues in distance metric learning. Within each of the four categories, we have summarized existing work, disclosed their essential connections, strengths and weaknesses. The first category is supervised distance metric learning, which contains supervised global distance metric learning, local adaptive supervised distance metric learning, Neighborhood Component Analysis (NCA) [13], and Relevant Components Analysis (RCA) [1]. The second category is unsupervised distance metric learning, covering linear (Principal Component Analysis (PCA) [14], Multidimensional Scaling (MDS) [5]) and nonlinear embedding methods (ISOMAP [35], Locally Linear Embedding (LLE) [30], and Laplacian Eigenamp (LE) [2]). We further unify these algorithms into a common framework based on the embedding computation. The third category, which is maximum margin based distance metric learning approaches, includes the large margin nearest neighbor based distance metric learning methods and semi-definite Programming (SDP) methods to solve the kernelized margin maximization problem. And the fourth category discussing kernel methods towards learning distance metrics, covers kernel alignment [28] and its SDP approaches [26], and also the extension work of learning the idealized kernel [25].

In addition to this survey [41], here we provide a complete and updated summarization of the related work on both unsupervised distance metric learning and supervised distance metric learning, including the most recent work in the area of distance metric learning.

Many unsupervised distance metric learning algorithms are essentially for the purpose of unsupervised dimensionality reduction, i.e. learning a low-dimensional embedding of the original feature space. This group of methods can be divided into nonlinear and linear methods. The well known algorithms for nonlinear unsupervised dimensionality reduction are ISOMAP [35], Locally Linear Embedding (LLE) [30], and Laplacian Eigenamp (LE) [2]. ISOMAP seeks the subspace that best preserves the geodesic distances between any two data points, while LLE and LE focus on the preservation of local neighbor structure. An improved and stable version of LLE is achieved in [45], by introducing multiple linearly independent local weight vectors for each neighborhood. Among the linear methods, Principal Component Analysis (PCA) [14] finds the subspace that best preserves the variance of the data; Multidimensional Scaling (MDS) [5] finds the low-rank projection that best preserves the inter-point distance given by the pairwise distance matrix; Independent components analysis (ICA) [4] seeks a linear transformation to coordinates in which the data are maximally statistically independent. Locality Preserving Projections (LPP) [17] and Neighborhood Preserving Embedding (NPE) [18] are the linear approximation to LE and LLE, respectively. Note that although LPP and NPE are developed originally for the unsupervised dimensionality reduction, they can also be extend to the setup of supervised distance metric learning where the label information is used to construct the weight matrix. In addition, several efforts have been made to achieve good generalization properties of manifold learning, i.e., the robustness to noise and nonlinear transformation. Locally Smooth Manifold Learning (LSML)[22] is able to recover a manifold from noisy data using weighted local linear smoothing, and effectively handle outliers. As an extension of LSML,[9] learns a representation of non-isometric and nonlinear manifold, which enables the manipulation of new-coming data points, under the concept of "generalization beyond the training data". In the effort of discovering low dimensional representations through constructing a semidefinite programs (SDPs) with low rank solutions, differing from previous approaches, [39] conducts matrix factorization that respects local distance constraints, and yields smaller SDPs than previous work and achieves good approximations to the original problem.

In this study, we limited ourselves to supervised distance metric learning, i.e. learning a distance metric from side information that is typically presented in a set of pairwise constraints. The optimal distance metric is found by keeping objects in equivalence constraints close, and at the same time, objects in inequivalence constraints well separated. In the past, a number of algorithms have been developed for supervised distance metric learning. [40] formulates distance metric learning into a constrained convex programming problem by minimizing the distance between the data points in the same classes under the constraint that the data points from different classes are well separated. This algorithm is extended to the nonlinear case in [25] by the introduction of kernels. Local linear discriminative analysis [16] estimates a local distance metric using the local linear discriminant analysis. Relevant Components Analysis (RCA) [1] learns a global linear transformation from the equivalence constraints. The learned linear transformation can be used directly to compute distance between any two examples. Discriminative Component Analysis (DCA) and Kernel DCA [20] improve RCA by exploring negative constraints and aiming to capture nonlinear relationships using contextual information. Essentially, Relevant Components Analysis (RCA) [1] and Discriminative Component Analysis (DCA) [20] can be viewed as extensions of Linear Discriminant Analysis (LDA) [10] by exploiting the must-link constraints and cannot-link constraints. Local Fisher Discriminant Analysis (LFDA) [34] extends LDA by assigning greater weights to those connecting examples that are nearby rather than distant. [23] provides an efficient incremental learning method for LDA, by adopting sufficient spanning set approximation for each update step. [33] extends the support vector machine to distance metric learning by encoding the pairwise constraints into a set of linear inequalities. Neighborhood Component Analysis (NCA) [13] learns a distance metric by extending the nearest neighbor classifier. The maximum-margin nearest neighbor (LMNN) classifier [38] extends NCA through a maximum margin framework. [12] learns a Mahalanobis distance that tries to collapse examples in the same class to a single point, and in the meantime keep examples from different classes far away. Local Distance Metric (LDM) algorithm [43] addresses multimodal data distributions in distance metric learning by optimizing local compactness and local separability in a probabilistic framework. [42] estimates a posterior distribution for the estimated distance metric through a full bayesian treatment; and actively selects unlabeled example pairs for labeling with the greatest uncertainty in relative distance. [32] learns a distance metric from relative comparison through SVM-like convex optimization. Locally Linear Metric Adaptation (LLMA) [15] presents a semi-supervised clustering approach that performs nonlinear transformation globally but linear transformation locally. [8] presents an LDA based approach as an efficient eigen problem. Previous study [38] has shown the LMNN algorithm delivered the state-of-the-art performance among all the distance metric learning algorithms.

A group of most recent work focuses on examining and exploring the relationship among metric learning, dimensionality reduction, kernel learning, and semi-supservied learning. [36] unifies the goals of dimensionality reduction and metric learning, in order to reduce the risks of overfitting. [21] studies the connections between statistical translation, heat kernels on manifolds and graphs, and expected distances, specifically for high dimensional structured data. [27] studies the connection between distance metric learning in a transductive framework and nonlinear dimensionality reduction.

Several information-theoretic approaches towards distance learning have been recently proposed, in addition to traditional distance metric learning that assumes a quadratic form for the distance between any two vectors. [7] expresses the learning a Mahalanobis distance function as a Bregman optimization problem, by minimizing the differential relative entropy (the LogDet divergence) between two multivariate Gaussians subject to linear constraints on the distance function. [19] defines the similarity as the gain in coding length by shifting from pairwise independent encoding to joint encoding. It has been shown in [19] that, in a certain large sample limit, coding similarity converges to the Mahalanobis metric estimated by the Relevant Components Analysis algorithm (RCA) [1].

There are some interesting pattern recognition literature on distance metric learning [24, 44, 6, 37, 29, 31]. [24] essentially is a metric learning method that implements a nonlinear mapping function by optimizing the parameters under a separability criterion. [44] propose a parametric learning method that finds a regression mapping of the input space, through which between-class dissimilarity is always larger than within-class dissimilarity. In [44], parameters are learnt iteratively by the majorization algorithm.

In the context of visual recognition, the recent development of distance learning includes the following. [11] learns a local perceptual distance function for each training example by combining elementary distances in the patch level. [3] measures similarity between two signals by composition, i.e. how easy it is to compose one signal from few large contiguous chunks of another.

As a summary to the above related work, Table and Table illustrate the key related approaches to unsupervised distance metric learning and supervised distance metric learning, respectively. Both Table and Table specify a few general properties for distance metric learning (for instance, linear or nonlinear, global or local) and describe the learning strategies.

References

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proc. Int. Conf. on Mach. Learn.*, 2003.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 2003.

[3] Oren Boiman and Michal Irani. Similarity by composition. In *NIPS*, pages 177–184. 2007.

[4] P. Common. Independent component analysis — a new concept? *Signal Processing*, 36:287–314, 1994.

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

[6] T. F. Cox and G. Ferry. Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, 26(1):145–153, 1993.

[7] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proc. Int. Conf. on Machine Learning*, pages 209–216, 2007.

[8] Cristianini N. De Bie T., Momma M. Efficiently learning the metric using side-information. In *Proc. Int. Conf. on Algorithmic Learning Theory*, 2003.

[9] Piotr Dollár, Vincent Rabaud, and Serge Belongie. Non-isometric manifold learning: analysis and an algorithm. In *Proc. Int. Conf. on Machine Learning*, pages 241–248, 2007.

[10] R. A. Fisher. The use of multiple measurements in taxonomic problems. In *Annual of Eugenics*, 1936.

[11] Andrea Frome, Yoram Singer, and Jitendra Malik. Image retrieval and classification using local distance functions. In *NIPS*, pages 417–424. 2007.

[12] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In *NIPS*, 2006.

[13] Jacob Goldberger, Sam Roweis, Geoffrey Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2005.

[14] R. Gonzales and R. Woods. *Digital Image Processing*. Addison-Wesley, 1992.

[15] D.Y. Yeung H. Chang. Locally linear metric adaptation for semi-supervised clustering. In *Proc. Int. Conf. on Machine Learning*, 2004.

[16] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6), 1996.

[17] X. He and Partha Niyogi. Locality preserving projections. In *NIPS*, 2003.

[18] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *Proc. Int. Conf. on Computer Vision*, 2005.

[19] Aharon Bar Hillel and Daphna Weinshall. Learning distance function by coding similarity. In *Proc. Int. Conf. on Machine Learning*, pages 65–72, New York, NY, USA, 2007.

[20] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proc. Computer Vision and Pattern Recognition*, 2006.

[21] G. Lebanon J. Dillon, Y. Mao and J. Zhang. Statistical translation, heat kernels, and expected distance. In *Proc. Uncertainty in Artificial Intelligence*, 2007.

[22] Hongyuan Zha JinHyeong Park, Z. Zhang and R. Kasturi. Local smoothing for manifold learning. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 452–459, 2004.

[23] T.-K. Kim, S.-F. Wong, B. Stenger, J. Kittler, and R. Cipolla. Incremental linear discriminant analysis using sufficient spanning set approximations. In *Proc. Computer Vision and Pattern Recognition*, 2007.

[24] W. L. G. Koontz and K. Fukunaga. A nonlinear feature extraction algorithm using distance information. *IEEE Trans. Computers*, 21(1):56–63, 1972.

[25] James T. Kwok and Ivor W. Tsang. Learning with idealized kernels. *Proc. ICML 2003*, 2003.

[26] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semi-definite programming. Technical Report UCB/CSD-02-1206, EECS Department, University of California, Berkeley, Oct 2002.

[27] Fuxin Li, Jian Yang, and Jue Wang. A transductive framework of distance metric learning by spectral dimensionality reduction. In *Proc. Int. Conf. on Machine Learning*, pages 513–520, 2007.

[28] A. Elisseeff N. Cristianini, J. Shawe-Taylor and J. Kandola. On kernel-target alignment. In *NIPS*, 2002.

[29] Y. LeCun P. Y. Simard and J. Decker. Efficient pattern recognition using a new transformation distance. In *NIPS*, volume 6, page 5058, 1993.

[30] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. In *Science*, 2000.

[31] R. Hadsell S. Chopra and Y. LeCun. Learning a similiarty metric discriminatively, with application to face verification. In *Proc. Computer Vision and Pattern Recognition*, 2005.

[32] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2003.

[33] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2004.

[34] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proc. Int. Conf. on Machine Learning*, 2006.

[35] J.B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2000.

[36] Lorenzo Torresani and Kuang chih Lee. Large margin component analysis. In *NIPS*, pages 1385–1392. 2007.

[37] A. R. Webb. Multidimensional scaling by iterative majorization using radial basis func- tions. *Pattern Recognition*, 28(5):753–759, 1995.

[38] Kilian Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[39] Kilian Q. Weinberger, Fei Sha, Qihui Zhu, and Lawrence K. Saul. Graph laplacian regularization for large-scale semidefinite programming. In *NIPS*, pages 1489–1496. 2007.

[40] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, 2003.

[41] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey, 2006. `http://www.cse.msu.edu/~yangliu1/frame_survey_v2.pdf`.

[42] Liu Yang, Rong Jin, and Rahul Sukthankar. Bayesian active distance metric learning. In *Proc. Uncertainty in Artificial Intelligence*, 2007.

[43] Liu Yang, Rong Jin, Rahul Sukthankar, and Yi Liu. An efficient algorithm for local distance metric learning. In *Proc. AAAI*, 2006.

[44] Z. Zhang, J. Kwok, and D. Yeung. Parametric distance metric learning with label information. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 2003.

[45] Zhenyue Zhang and Jing Wang. Mlle: Modified locally linear embedding using multiple weights. In *NIPS*, pages 1593–1600. 2007.

Methods	Properties
Principal Component Analysis (PCA) [14]	global structure preserved, linear, best preserve the variance of the data
Multidimensional Scaling (MDS) [5]	global structure preserved, linear, best preserve inter-point distance in low-rank
Independent Components Analysis (ICA) [4]	global structure preserved, linear, transformed data are maximally statistically independent
Locality Preserving Projections (LPP) [17]	local structure preserved, linear, approximation to LE
Neighborhood Preserving Embedding (NPE) [18]	local structure preserved, linear, approximation to LLE
ISOMAP [35]	global structure preserved, nonlinear, glopreserve the geodesic distances
Locally Linear Embedding (LLE) [30]	local structure preserved, nonlinear, preserve local neighbor structure
Laplacian Eigenamp (LE) [2]	local structure preserved, nonlinear, preserve local neighbor structure
Locally Smooth Manifold Learning (LSML) [22]	local structure preserved, nonlinear, manifold recovery by weighted local linear smoothing

Table 1: Unsupervised distance metric learning methods. This group of methods essentially learn a low-dimensional embedding of the original feature space; and can be categorized along two dimensions: preserving glocal structure vs. preserving local structure; and linear vs. nonlinear

Methods	Properties
Local Linear Discriminative Analysis [16]	local, linear, estimate a local distance metric using the local linear discriminant analysis
Global Distance Metric Learning [40]	global, linear, constrained convex programming problem
Relevant Components Analysis (RCA) [1]	global, linear, learn a global linear transformation from equivalence constraints
Discriminative Component Analysis (DCA) and Kernel DCA [20]	global, linear, improve RCA by exploring negative constraints
Local Fisher Discriminant Analysis (LFDA) [34]	local, linear, extend LDA by assigning greater weights to closer connecting examples
Neighborhood Component Analysis (NCA) [13]	local, linear, extend the nearest neighbor classifier to metric learning
Maximum-Margin Nearest Neighbor (LMNN) Classifier [38]	local, linear, extend NCA through a maximum margin framework
Localized Distance Metric Learning (LDM) [43]	local, linear, optimize local compactness and local separability in a probabilistic framework
Baysian Active Distance Metric Learning [42]	global, linear, select example pairs with the greatest uncertainty, posterior estimation with a full Bayesian treatment

Table 2: Supervised distance metric learning methods