SEMI-SUPERVISED NONLINEAR DISTANCE METRIC LEARNING VIA RANDOM FOREST AND RELATIVE SIMILARITY ALGORITHM

N. Saranya [1], C. Usha Nandhini[2]

[1] *Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India*
[2] *Assistant Professor, Dept. of Computer Applications, Vellalar College for Women, Erode, Tamilnadu, India*

---***---

***Abstract** - Similarity measure is closely related to distance metric learning. Metric learning is the task of learning a distance function over objects. In the base work, a nonlinear machine learning method is implemented by using Semi-Supervised Max-Margin Clustering to construct a forest of cluster hierarchies. In that individual component of the forest represent cluster hierarchies. Clustering hierarchies gives handling any form of similarity or distance. It is also used for applicability to any attributes type. Most hierarchal algorithms do not revisit once constructed clusters with the purpose of improvement. For distance metric learning give some computational complexity. To reduce the complexity and improvement purpose, proposed algorithm called Relative Similarity use the linear reconstruction weights to measure the similarity between the adjacent points. The original data points are collected in dimensional space and the goal of the algorithm is to reduce the dimensionality. The proposed algorithm gives good clustering results and reasonably fast for sparse data sets of several thousand elements. The investigational tests prove that the leads to enhanced performance than the presented approach by analyzing the code quality in an efficient manner.*

***Key Words:** Semi-Supervised, Distance Metric learning, Relative Similarity algorithm, Max-Margin clustering, Locally Linear Embedding, Random Forest.*

1. INTRODUCTION

Data mining problems—nearest neighbor classification, retrieval, clustering—are at their core dependent on the availability of an effective measure of pairwise distance. Ad hoc selection of a metric, whether by relying on a typical such as Euclidean distance or attempt to select a domain appropriate kernel, is unreliable and inflexible. It is thus attractive to advance metric selection as a learning problem, and attempt to train strong problem-specific distance measures using data and semantic in a row. A wide range of methods have been proposed to address this learning problem, but the field has traditionally been conquered by algorithms that assume a linear model of distance, particularly Mahalanobis metrics. This attitude provides two significant contributions: first, unlike previous tree-based nonlinear metrics, it is semi-supervised, and can incorporate information from both controlled and unconstrained points into the learning algorithm. This is an important pro in many problem settings, mainly when scaling to larger datasets where only a tiny quantity of the full pair wise constraint set can realistically is unruffled or used in training. Second, the iterative, hierarchical nature of our training process allows relaxing the constraint satisfaction problem. Rather than attempting to satisfy every accessible constraint simultaneously, at each hierarchy node we can optimize an appropriate constraint subset to focus on, parting others to be addressed lower in the tree (or in other hierarchies in the forest). By select constraints in this way, can avoid situation where attempting to satisfy confused constraints, and thereby better model hierarchical data structures. Semi-supervised learning is a unification of supervised and unsupervised learning. It incorporates the information's from both constrained and unconstrained points. k-Nearest Neighbor classifier uses a metric to identify the nearest neighbors and many clustering algorithms. The mean is to learn a function which is able to make straightforward the well on invisible data. Semi-supervised learning models include self-training, mixture models, graph-based methods, co-training and multi view learning.

2. LITERATURE REVIEW

David M. Johnson, Caiming Xiong, and Jason J. Corso[4] Semi-supervised max-margin clustering to construct a forest of cluster hierarchies, where each individual hierarchy can be interpreted as a weak metric over the data. These experiments were conducted on the k-nearest neighbor classification task, with k ¼ 11, using three-fold cross-validation. In each case 1 percent of all must-link and 1 percent of all cannot-link pairwise constraints were used. On the USPS set, the two techniques yield essentially identical results. While high-dimensional data (and particularly data where d n) presents a challenge to any machine learning application, it is particularly troublesome for traditional Mahalanobis metrics. Solving for the Mahalanobis matrix M requires optimizing d2 independent variables. When d is large, this quickly becomes both prohibitively costly and analytically dubious. By contrast, HFD needs only a subset of the available features in each node, and computes only a linear combination over this subset.

Bellet, and Amaury Habrard [1] proposed an appropriate ways to calculate the distance or similarity between data is ubiquitous in machine learning, pattern recognition and data mining, but handcrafting such good metrics for definite problems is usually difficult. This has led to the appearance of metric learning, which aims at automatically learning a metric from data and has attracted a lot of interest in machine learning and related fields for the past ten years. This proposes a systematic review of the metric learning literature, highlighting the pros and cons of each approach a well-studied and successful framework, but additionally presents a wide range of methods that have recently emerged as powerful alternatives, including nonlinear metric learning, similarity learning and local metric learning. Nevertheless, recent advances such as GESL have shown that drawing inspiration from successful feature vector formulations (even if it requires simplifying the metric) can be highly beneficial in terms of scalability and flexibility.

Brian J. V. Davis Kulis, Prateek Jain PJAIN,Suvrit Sra and Inderjit S. Dhillon [3] proposed information-theoretic approach to learning a Mahalanobis distance function. This work formulates the problem as that of minimizing the degree of difference relative entropy between two multivariate Gaussians under constraints on the distance function. These works express this problem as a particular Bregman optimization problem—that of minimizing the LogDet deviation subject to linear constraints. Note that both MCML and LMNN are not amenable to optimization subject to pairwise distance constraints. Instead, it will compare method to the semi-supervised clustering algorithm HMRF-KMeans.

Kedem.D, Tyree.S, Weinberger.K, Sha.F, and Lanckriet.G[5] analyzed to develop a new framework of kernelizing Mahalanobis distance learners. The new KPCA trick framework offers several realistic compensations over the classical kernel trick framework, e.g. no mathematical formulas and no reprogramming are required for a kernel implementation, a way to speed up an algorithm is provided with no extra work, and the framework avoids troublesome problems such as singularity. Advantages of our framework over the classical kernel-trick framework have been illustrate and evidence showing satisfiable show of kernel arrangement has been reported on three recent algorithms which previously did not have kernel versions. This work can be complete to pattern recognition in other settings, and this chance will be the main subject of our future work.

Leo Breiman[6] have proposed a Random Forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The simplification error for forests converges to a limit as the number of trees in the forest becomes large. Internal estimates monitor error, potency, and correlation and these are used to show the response to growing the number of features used in the splitting. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation. For a while, the conventional philosophy was that forests could not struggle with arcing type algorithms in terms of accuracy. Their accuracy indicates that they act to reduce bias. The mechanism for this is not obvious. A recent work shows that in distribution space for two class problems, random forests are equivalent to a kernel acting on the true margin. The theoretical framework given by Kleinberg for Stochastic Discrimination may also help sympathetic.

3. PROBLEM FORMULATION

Clustering algorithm generally used for grouping the data among the relationship of the two data points. In the existing work Semi-Supervised Max-Margin clustering is used for distance measure. In the Proposed work Relative Similarity algorithm is used for better accuracy comparing to the previous algorithm. A relative similarity and dissimilarity matrix deduced from the new similarity measure was then outline to demonstrate the similarity of different sequences.

3.1 Existing Scenario

In the existing system non linear metric learning method is proposed using semi-supervised max-margin clustering to construct a forest of cluster hierarchies. In this individual hierarchy is produced as weak metric. To get powerful and robust nonlinear metric model introduce randomness during hierarchy training. In HFD model, it is composed of number of trees and trained independently and splitting function is in linear combination. In that individual component of the forest represent cluster hierarchies. HFD requires a robust approach to the hierarchy splitting problem that reliably generates semantically meaningful splits. Additionally, in order to allow for efficient metric inference, splitting algorithm must generate explicit and efficiently evaluable splitting functions at each node. This can be attaining by using max margin clustering. In this novel in-metric approximate nearest-neighbor retrieval algorithm is proposed for method that greatly decreases retrieval times for large data with little reduction in accuracy.

3.2 Proposed Scenario

The future algorithm uses linear reconstruction weights to measure the similarity between adjacent points. Then based on the constructed connected graph, the new path-based similarity can be got. In this algorithm use the linear reconstruction weights to measure the similarity between the adjacent points. Call it as relative similarity. In LLE, the linear reconstruction weights can be negative. First build a Minimum Spanning Tree (MST) on the defined graph (the time complexity is O(n2log(n))), then use. Then finally get clustering results of path-based relative similarity clustering.

Advantage of Proposed Scenario

- It increases the semi-supervised max-margin clustering accuracy in superior.
- It reduces the computational complexity.
- Some basic metric to extract the most descriptive terms in a document.
- Easily compute the similarity between two documents using it.

4. SYSTEM METHODOLOGY

- Semi-Supervised Max-Margin hierarchy Forests
- HFD learning and interface
- LLE algorithm
- Random forest
- Relative similarity algorithm

Semi-Supervised Max-Margin Hierarchy Forests

In this section describe in detail Hierarchy Forest Distance (HFD) model, as well as the procedures for training and inference. HFD does not fit the strict definition of a distance metric.

Hierarchy Forests

The structure of the HFD model draws some basic elements from random forests, in that it is composed of T trees trained independently in a semi-random fashion, with individual nodes in the trees defined by a splitting function that divides the local space into two or more segments. HFD is conceptually distinct from random forests in that the individual components of the forest represent cluster hierarchies rather than decision trees. HFD also differs from the most common form of random forest in that it does not do bootstrap sampling on its training points, and its splitting functions are linear combinations rather than single-feature thresholds.

$$\mathrm{D(a,b)}=1/T \sum_{t=1}^{T} H_t(a,b)$$

Additionally, in order to allow for efficient metric inference, splitting algorithm must generate explicit and efficiently evaluable splitting functions at each node. Given these constraints, advance the hierarchy learning problem as a series of increasingly fine-grained flat semi-supervised clustering problems, and solve these flat there are significant differences between the two methods. HFD is conceptually distinct from random forests (and the Random Forest Distance (RFD) metric.

HFD Learning and Interference

The fact the trees used in HFD represent cluster hierarchies rather than decision tree significant implications for HFD training, imposing stricter requirements on the learned splitting functions. While the goal of decision tree learning is finally to defer a set of pure single-class leaf nodes, a cluster hierarchy instead seeks to accurately group data elements at every stage of the tree. Thus, if the hierarchy learning algorithm divides the data poorly at or near the root node, there is no way for it to recover from this error later on. This is moderately mitigated by learning a forest in place of a single tree, but even in this case the common of hierarchies in the forest must correctly model the high-level semantic relationship between any two data elements. For this reason, HFD requires a robust approach to the hierarchy splitting problem that reliably generates semantically meaningful splits.

$$H_t\mathrm{(a,b)}=\begin{cases} 0 & if\ H_{tl}(a,b) is\ aleaf\ node \\ P_t(a,b).\frac{|H_{tl}(a,b)|}{N} & otherwise \end{cases}$$

Clustering problems via max-margin clustering Max-margin clustering has a number of advantages that make it ideal for problem. In addition to their widespread use in support vector machines for classification, max-margin and large-margin techniques have proven highly effective in the metric learning domain and many, including MMC, can be solved in linear time.

$$P_t(a,b) = \frac{1}{1+\exp(\propto.p_{tl(a,b)}(X_a))} - \frac{1}{1+\exp(\propto.p_{tl(a,b)}(X_b))}$$

Most importantly, MMC returns a simple and explicit splitting function which can be computed efficiently and applied to points outside the initial clustering. A relaxed form semi-supervised MMC (SSMMC). This uses pair wise must-link (ML) and cannot-link (CL) constraint to recover semantic clustering performance. Constraints of this type specify either semantic similarity (ML) or dissimilarity (CL) between pairs of points, and do not require the availability of class labels.

LLE Algorithm

Suppose the data consist of N real-valued vectors each of dimensionality D, sampled from some underlying manifold. Expect each data point and its neighbors to recline on or close to a locally linear patch of the multiple. Characterize the local geometry of these patches by linear coefficients that reconstruct each data point from its neighbors. Reconstruction errors are measured by the cost function

$$\varepsilon(W) = \sum_i \left| \underset{X_i}{\rightarrow} - \sum_j w_{ij} \underset{X_j}{\rightarrow} \right|$$

This adds up the squared distances between all the data points and their reconstructions. Minimize the cost function subject to two constraints:

First, all data point$\underset{x_i}{\rightarrow}$ is reconstructed only from its neighbors, enforcing $W_{ij} = 0 \; if \; \underset{X_i}{\rightarrow}$ do not be in the right place to the set of neighbors of $\underset{x_i}{\rightarrow}$

- Second that the rows of the weight matrix sum to one $\sum_j w_{ij} = 1$
- Each high-dimensional observation $\underset{Xi}{\rightarrow}$ is mapped to a low-dimensional vector $\underset{Y_i}{\rightarrow}$ representing global internal coordinates on the manifold.
- This is done by choosing d-dimensional coordinate $\underset{Y_i}{\rightarrow}$to minimize the embedding cost function

$$\varphi(y) = \sum_i \left| \underset{Y_i}{\rightarrow} - \sum_j W_{ij} \underset{Y_j}{\rightarrow} \right|$$

- This cost function, like the previous one, is based on locally linear reconstruction errors, but here fix the weights W_{ij} while optimizing the coordinates$\underset{Yi}{\rightarrow}$.
- After that define similarity of data points. Similarity measure that quantified the comparison between two objects.

Although no single definition of a similarity measure exists, usually such measures are in some sense the inverse of distance metrics take on large values for similar objects and either zero or a negative value for very dissimilar objects. A minimum spanning tree (MST) or minimum weight spanning tree is a division of the edges of a connected, edge-weighted undirected graph that connect all the vertices together, without any cycles and with the minimum possible total edge weight.

Random Forest

Random Forests grows several classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and says the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree are grown as follows

- If the number of cases in the training set is N, sample N cases at random - but with alternate, from the original data. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are select at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
- Each tree is developed to the largest extent possible. There is no pruning.

Relative Similarity Algorithm

The main difference between new path-based spectral clustering and the existed path-based clustering is the definition of the path-based similarity. Inspired by Locally Linear Embedding (LLE) algorithm and Ref, can use the linear reconstruction weights to measure the the similarity among the adjacent points. Call it relative similarity. In LLE, the linear reconstruction weights can be negative. While in algorithm, force the weights must be nonnegative. So use the same method in Ref to compute the similarity between data points. The objective function is defined as below:

$$min \sum_i \left\| x_i - \sum_{j:x_j \in N_{(xi)}} w_{ij} x_j \right\|^2$$

$$s.t \; \sum_{j:x_j \in N_{(xi)}} w_{ij} = 1, \quad w_{ij} \geq 0$$

Where $N_{(xi)}$) denotes the set composed of p nearest neighbors of *xi*. Obviously, the more *xi* is similar to x_j , the closer w_{ij} approximates 1. On the opposite side, w_{ij} will approach to zero. Be sides the general situation $w_{ij} \neq w_{ji}$. It can do some deduction on the upper equation.

$$min \sum_i \left\| x_i - \sum_{j:x_j \in N_{(xi)}} w_{ij} x_j \right\|^2$$

$$min \sum_i \left\| \sum_{j:x_j \in N_{(xi)}} w_{ij} (x_i - x_j) \right\|^2$$

$$min \sum_i \sum_{j,k:x_j,x_k \in N_{(xi)}} w_{ij} G^i_{jk} w_{ik}$$

Where G^i_{jk} denotes the (*l, j*) element of the Gram matrix G^i_{jk} = $(x_i - x_j)^T (x_i - x_k)$. So the reconstruction weights can be get through the below *n* quadratic programming problems, namely

$$minw_{ij} \sum_{j,k:x_j,x_k \in N_{(xi)}} w_{ij} G^i_{jk} w_{ik}$$

$$s.t. \sum_j w_{ij} = 1, \;\; w_{ij} \geq 0$$

Equation 10, measure the similarity of data points. As $w_{ij} \neq w_{ji}$., redefine the relative similarity as $s_{ij} = s_{ji} = w_{ij} + w_{ji/}2$.Then the adjacent matrix W' of the data points can be got,

$$(S)_{i,j} \begin{cases} s_{ij}, & x_j \epsilon N_{(x_i)} or \; x_i \, \epsilon N_{(x_j)} \\ & 0, otherwise \end{cases}$$

Then through the defined matrix W', can construct a connect graph for some data set. According to the definition of path-based similarity, can get the path-based relative similarity.

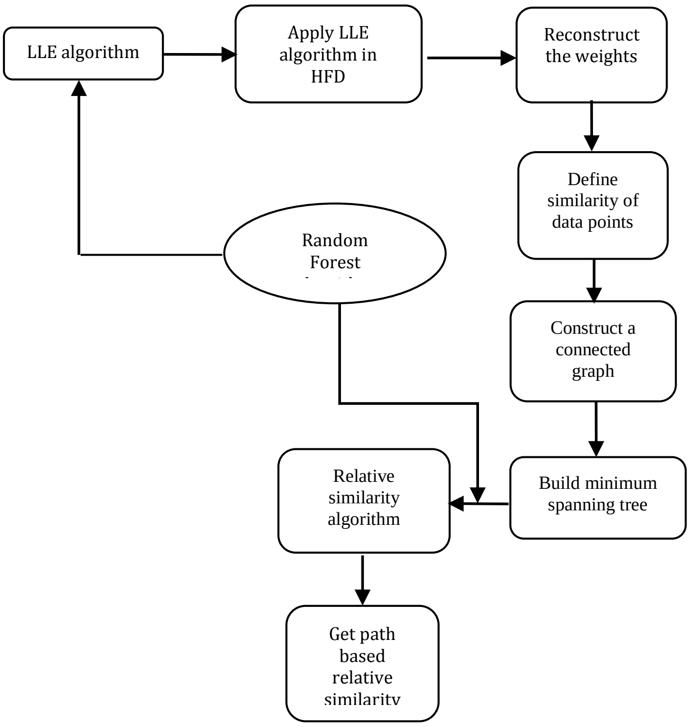
LLE algorithm
Apply LLE algorithm in HFD
Reconstruct the weights
Define similarity of data points
Random Forest
Construct a connected graph
Build minimum spanning tree
Relative similarity algorithm
Get path based relative similarity

Fig-1: System Architecture for RS

5. RESULTS AND DISCUSSION

Experimental Result

This section presents the experimental results that are performed to prove the proposed semi-supervised max-margin clustering system achieves high accuracy. The performance of the proposed semi-supervised max-margin clustering system is evaluated in terms of precision, recall, accuracy and f-measure with existing classification system.

DataSet Collection

The breast cancer diseases were predicted through RS algorithm. The dataset collected from the UCI Machine Learning Repository. According to the clustering scheme a particular disease was detected using breast cancer dataset. Dataset consist of 768 instances and eight attributes. All attribute values are only Numerical. The attributes are Number of times pregnant, Plasma Glucose concentration a 2 hour in an oral Glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2-hour sereem insulin, Body Mass Index, Diabetes pedigree functions, age, class variable(0 or 1).

Accuracy

The accuracy percentage of true results (both true positives and true negatives) among the total number of cases examines. Accuracy refers to the proximity of an exact value to a standard or known value.

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative+False positive+ False negative}}$$

Precision

Precision value is evaluated according to the feature classification at true positive false positive prediction. Precision is a description of random errors, a measure of statistical variability. It is expressed as follows

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

Recall

Recall value is evaluated according to the feature classification at true positive prediction, false negative. Recall in memory refers to the mental process of retrieval of information from the past. It is given as,

$$Recall = \frac{Truepositive}{(Truepositive + Falsenegative)}$$

F-Measure

F-measure is calculated from the precision and recall value. Precision is also used with recall, the percent of all relevant documents that is return by the search. The two measures are sometimes used together in the F1 Score to provide a single measurement for a system. It is calculated as,

F-measure = $2 \times (precision \times recall/precision + recall)$

Table-1 Comparison Table

Metrics	Semi-Supervised Max Margin Clustering	Relative Similarity
Accuracy	87.1866	94.9333
Precision	0.8516	0.9340
Recall	0.9080	0.9620
F-Measure	0.8789	0.9478

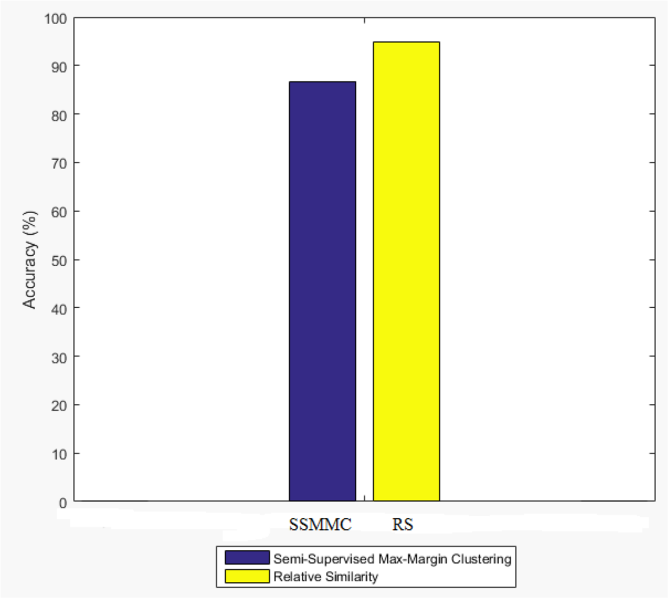

Chart-1: Accuracy Table

6. CONCLUSION AND FUTURE WORK

Semi-supervised nonlinear distance metric learning procedure based on forests of cluster hierarchies constructed via an iterative max margin clustering procedure. A non linear machine learning method uses semi supervised max-margin clustering with Relative Similarities to construct a forest of cluster hierarchies. Proposed an algorithm called RS algorithm is give better results compared to existing system. In LLE, the linear reconstruction weights can be negative. The proposed semi-supervised max-margin clustering method is effective in finding informative patterns to represent the sequences, leading to classification accuracy that is in most cases higher than the existing work.

The basic firefly algorithm is very efficient, that the solution is still shifting as the optima are approaching. It is possible to improve the solution quality by reducing the randomness gradually. A further improvement on the convergence of the algorithm is to vary the randomization parameter so that it decreases regularly as the optima are approaching. These could form important topics for further research. The Firefly Algorithm can be modified to solve multi objective optimization problems. In addition, the application of firefly algorithms in combination with other algorithms may form an exciting area for further research.

REFERENCES

[1] Bellet.A, Habrard.A, and Sebban.M, “A survey on metric learning for feature vectors and structured data,” arXiv preprintarXiv:1306.6709, 2013.

[2] Brian Kulis, “distance function and metric learning”, 2010

[3] Brian J. V. Davis Kulis, Prateek Jain PJAIN,Suvrit Sra and Inderjit S. Dhillon, “Information-theoretic metric learning”, pp-3306-3313, 2016.

[4] David M. Johnson, Caiming Xiong, and Jason J. Corso” Semi-Supervised Nonlinear Distance MetricLearning via Forests of Max-Margin Cluster Hierarchies” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL.28, 2016

[5] Kedem.D, Tyree.S, Weinberger.K, Sha.F, and Lanckriet.G,“Non-linear metric learning,” in Proc. Adv. Neural Inf. Process.Syst., 2012 .

[6] Leo Breiman, “Random Forests”,2001.

[7] Michael Perrot, and Amaury Habrard , “Regressive Virtual Metric Learning”, volume-25, 2016.

[8] Sumit Chopra, Raia Hadsell and Yann LeCun “Learning a Similarity Metric Discriminatively, with Application to Face verification” pp-539-546, 2005.

[9] Xiong.C, Johnson.D.M, Xu.R, and Corso.J.J, “Random forests for metric learning with implicit pairwise position dependence,”in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,2012.

[10] X Zhu, Andrew B. Goldberg,” Introduction to Semi-Supervised Learning Synthesis Lectures on Artificial Intelligence and Machine Learning” University of Wisconsin, Madison2009, 130 pages.

[11] Ying.Y and Li.P, “Distance metric learning with eigenvalue optimization,”J. Mach. Learn. Res., vol. 13, pp. 1–26, 2012.