

# Learning Local Semantic Distances with Limited Supervision

Shiyu Chang\*, Charu Aggarwal†, Thomas Huang\*

\* Beckman Institute, University of Illinois at Urbana-Champaign, IL 61801.

† IBM T.J. Watson Research Center, NY, 10598.

Email: {chang87, t-huang1}@illinois.edu, charu@us.ibm.com

**Abstract**—Recent advances in distance function learning have demonstrated that learning a good distance metric can greatly improve the performance in a wide variety of tasks in data mining and web search. A major problem in such scenarios is the limited labeled knowledge available for learning the user intentions. Furthermore, distances are inherently local, where a single global distance function may not capture the distance structure well. A challenge here is that local distance learning is even harder when the labeled information available is limited, because the distance function varies with data locality. To address these issues, we propose a local metric learning algorithm termed Local Semantic Sensing (LSS), which augments the small amount of labeled data with unlabeled data in order to learn the semantic information in the manifold structure, and then integrated with supervised intentional knowledge in a local way. We present results in a retrieval application, which show that the approach significantly outperforms other state-of-the-art methods in the literature.

## I. INTRODUCTION

The problem of distance function design is well known in the literature, and is an important task in the context of important data mining and web retrieval tasks. Distance function design is often used as a subroutine in these applications, and the quality of the final results are often dependent upon the underlying distance function. For example, information retrieval engines utilize the learned distance metric to measure the relevance of the candidate data to a query [20][21], or a similar pattern in the context of pattern retrieval [14]. Applications of this approach also exist in the context of clinical decision support, search, and retrieval settings [25].

In recent years, many sites such as *like.com*<sup>1</sup> use content-based retrieval, where the target of the search is an image or other complex object, rather than a set of keywords. For such applications, the distance between two objects is highly impacted by the *intention* of the end-user, because multiple ways exist for measuring similarities between complex objects. For such high dimensional domains representing complex objects, a significant *semantic gap* exists between the user intentions during retrieval and the results retrieved by using an unsupervised distance function. In recent years, many studies have demonstrated either theoretically or empirically, that user supervision can greatly improve the performance in retrieval and other data mining applications. However, most of these existing methods are still not very effective, especially in the context of retrieval tasks. This is because of two main problems, which feed into each other:

- *Local Nature of Distance Functions*: Distance functions are inherently *local*, where the structure of the distance function varies significantly with the data locality. This is a significant problem for many search and retrieval applications. For example, in a clinical application, the structure of the distance function associated with the search on a data record with the clinical characteristics of a cancer patient is likely to be very different from a search on a data record with the clinical characteristics of an AIDS patient.
- *Limited supervision*: Distance function learning heavily relies on the amount of supervised information. The requirements are even greater, when the distance function needs to be learned locally. However, it is not reasonable to expect that a given user can manually provide large amounts of feedback. This can result in over-fitting problems, especially when the underlying data is of very high dimensionality.

It should be pointed out, that while labeled data is difficult to collect, a significant amount of unlabeled data is often available in many real scenarios. The unlabeled data is useful, because it allows the learning of the manifold structure of the underlying high-dimensional data. It should be pointed out that data points which are contiguously located along manifolds are often semantically related to one another, though the structure of the semantic distances is often local to the specific location on the manifold (or query point). Therefore, if unsupervised data can be used in order to learn the semantic structure of the data in terms of the underlying manifolds, the supervision can be used primarily to learn the local variation of the distances between data points along the manifold in an application-specific way. As we will see, the latter can be learned quite effectively with a limited amount of data, as long as (the copious amount of) unsupervised data is used to learn the overall semantic structure.

In order to illustrate the advantages of a local approach such as the one presented in this work, consider the simple two moon synthetic example shown in figure 1 embedded in a 30-dimensional space, where the first two dimensions are illustrated, and the remaining are ambient noise. Figure 1(a) illustrates the local distance function contour of the Euclidean distance, which is invariant with data locality. Figure 1(b) shows the distance function contour, which varies significantly with the data locality, even without supervision, because of the changes in the manifold (and corresponding semantic) structure of the underlying data. Furthermore, local variations in the labeled data, can impact this basic contour even further,

<sup>1</sup><http://www.crunchbase.com/company/like>

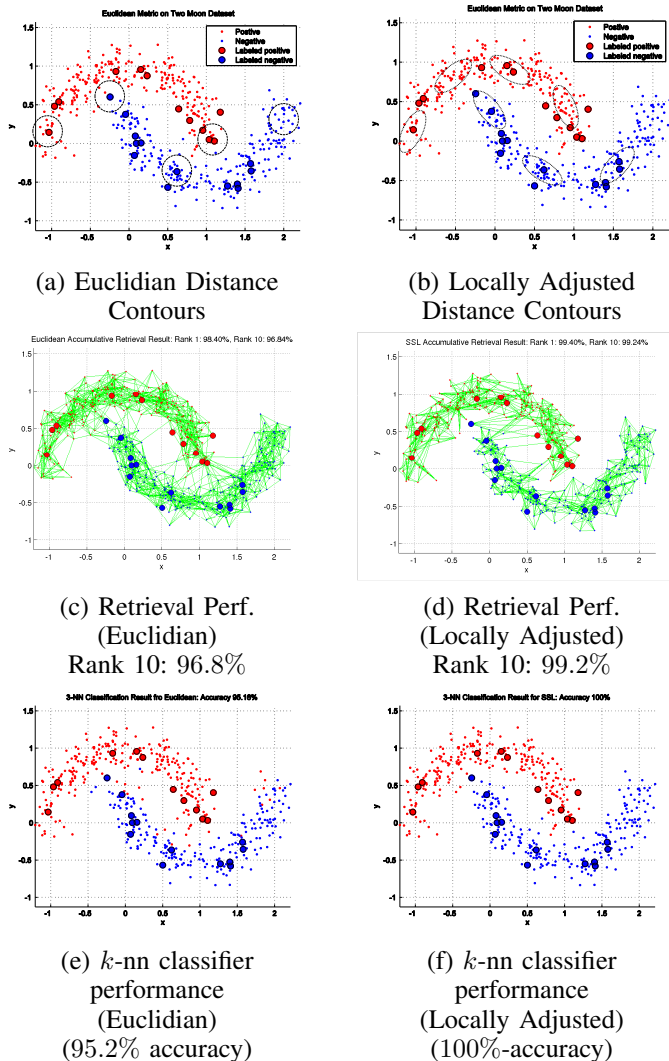


Fig. 1. Motivation and advantages of the proposed SSL Metric over traditional Euclidean measure. It is evident that the proposed approach is superior in terms ranking quality and classification.

though the amount of (labeled) data required to do this would be much less than would be required to do so from scratch. It is clear from the illustration, that the *local* distortion of this distance contour significantly helps in bringing truly similar points together. In Figures 1(c)–(f), we have shown some sample retrieval and classification quality results from the use of the global Euclidian distance, and a local approach (which in this case happens to be our local approach presented later in this work). It is evident that the local approach is superior in terms of ranking quality and the use of the distance function for a nearest neighbor classifier. The difference in these retrieval results is simply because of the failure of the global approach to adjust effectively to the varying local structure of the data, both in terms of the semantic structure implied by the unsupervised instances, and the user-intensional knowledge implied by the labeled instances. As we will see in detail in the experimental section, such a local approach provides significant advantages over different kinds of distance functions used for retrieval in the literature. Of course, the challenge here is that the amount of labeled data required in this case is significantly greater. Correspondingly, we use

unlabeled data in order to assist the learning process.

We propose a novel approach within a semi-supervised learning framework called *Local Semantic Sensing (LSS)*. Unlike traditional global methods, our approach constructs instance-centric (or query-centric) distances by harnessing the latent semantic knowledge in the unsupervised data in conjunction with the limited intentional knowledge provided by the user. In essence, our approach is able to seamlessly bridge the *local heterogeneity gap*, the *semantic gap* and *intensional gap* for effective retrieval within this integrated framework. Therefore, each individual instance can “sense” its own local semantic structure within a neighborhood, and combine it with the limited intensional knowledge provided by the user.

The remainder of this paper is organized as follows. We present the mathematical model for the semi-supervised LSS framework in section II. We discuss the intrinsic manifold dimension and a possible out-of-sample extension of the proposed method in section III. In section IV, we present extensive experiments on a wide range of data sets. Section V reviews related work on metric and manifold learning. The conclusions and future research directions are presented in section VI.

## II. LOCAL SEMANTIC SENSING

In this section, we will present the problem definition, model, and the algorithm for local semantic distance function learning. We will start by presenting the preliminaries.

### A. Problem Definition and Preliminaries

We begin by introducing the basic notation and terminology used in this paper. Assume that the input set of data examples are denoted by  $\mathcal{X} = \{x_1, \dots, x_n\} \in \mathbb{R}^d$ . Traditional metric learning calculates the distance between data vectors  $x_i$  and  $x_j$  through the generalized Mahalanobis measure, for many high dimensional content-based retrieval applications such as image data sets. Such distances become relevant, when querying web image with the use of another target image object, as opposed to a keyword-based measure. This kind of retrieval has recently started gaining popularity, though our approach is fairly general, and could apply to a wide variety of scenarios. The squared Mahalanobis distance between two multidimensional objects is defined as follows:

$$d_M^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j), \quad (1)$$

where  $M$  is an arbitrary symmetric positive semi-definite matrix. An important special case is one in which  $M$  is set to the identity matrix  $I_d$ . In such a case, the distance measure in Eq.(1) becomes the standard Euclidean distance.

Eq. (1) can also be viewed from the perspective of applying a global projection on all data points. Since  $M$  is a symmetric positive semi-definite metric, we can decompose  $M$  as  $M = U\Lambda U^T$  via eigen-decomposition, where  $U$  contains all possible eigenvectors of  $M$ , and  $\Lambda$  is a diagonal matrix with eigenvalues of  $M$ . Then, the squared Mahalanobis distance can be expressed as follows:

$$\begin{aligned} d_M^2(x_i, x_j) &= (x_i - x_j)^T U\Lambda U^T (x_i - x_j) \\ &= (x_i - x_j)^T (U\Lambda^{\frac{1}{2}})(\Lambda^{\frac{1}{2}}U^T)(x_i - x_j) \\ &= (x_i - x_j)^T LL^T(x_i - x_j) \\ &= \|L^T(x_i - x_j)\|_2^2 \\ &= \|\tilde{x}_i - \tilde{x}_j\|_2^2, \end{aligned} \quad (2)$$

where  $\tilde{x} = L^T x$ . Therefore, the squared Mahalanobis distance between two data points can be either parameterized by a  $d \times d$  semi-positive definite matrix or a  $d \times p$  low-dimensional projection matrix  $L$ , where  $p < d$ . The generalized Mahalanobis distance is equivalent to the Euclidean distance in transformed space.

Furthermore, learning a proper metric requires a large amount of labeled information even in the case of a global metric. The supervision information is usually given as either (a) explicit label information; (b) pairwise constraints; (c) triplet setup; and (d) quarto partial ordering. The explicit label information is infeasible to obtain in most data mining and web search applications, while the pairwise constraint is easier to obtain. Unlike pairwise constraints, triplets and partial ordering are rarely used in ranking related applications, because the number of constraints increases with number of samples dramatically. Therefore, we assume supervised information is available in the form of pairwise constraints. Specifically, we model the supervised information in the form of positive and negative similarity comparisons among data samples:

$$\begin{aligned} \mathcal{S} &= \{(\mathbf{x}_i, x_j) : x_j \text{ is similar to } x_i\} \\ \mathcal{D} &= \{(\mathbf{x}_i, x_j) : x_j \text{ is not similar to } x_i\}. \end{aligned} \quad (3)$$

$\mathcal{S}$  denotes a set of positive similarity constraints, while  $\mathcal{D}$  denotes negative constraints. It is worth mentioning that pairwise constraints are asymmetrical, which differs from other traditional setups. This is a natural consequence of the locality in the distance computation, since the distance computation from  $x_i$  to  $x_j$  with the former as the query point, is local to  $x_i$ . On the other hand, the distance computation with  $x_j$  as the query point is local to  $x_j$ . Of course, such asymmetry may affect the retrieval results only slightly especially when the retrieved data points are very similar to the query points.

### B. Local Semantic Model

The problem of learning distance function  $d(\cdot, \cdot)$  from pairwise inputs can be seen as learning a mapping function  $f$  to a new feature space, such that a traditional distance between  $f(x_i)$  and  $f(x_j)$  in the new feature space, is same as that in the original feature space. With this definition, distance function learning can be categorized as *linear* or *nonlinear* on whether the  $f$  is linear or not. As we have seen in the toy example illustrated in figure 1, linear mapping functions are often unable to capture local discriminative information from a given data distribution. We formally introduce or LSS distance as follows:

*Definition 1 (Local Semantic Distance):* The local semantic distance is defined local to each instance  $x_i$ , with the use of a local distortion matrix  $M_i$ . The squared distance between two data instances  $x_i$  and  $x_j$  can be defined as follows:

$$\begin{aligned} d_{M_i}^2(x_i, x_j) &= (x_i - x_j)^T M_i (x_i - x_j) \\ d_{M_j}^2(x_j, x_i) &= (x_j - x_i)^T M_j (x_j - x_i). \end{aligned} \quad (4)$$

Note that the distortion matrices  $M_i$  are parameterized, and eventually learned by the approach. The first equation indicates the distance from *query point*  $x_i$  (first argument) to *data point*  $x_j$  (second argument). In the second case,  $x_j$  is the query point. Thus, this particular definition of distance is asymmetrical,

---

### Algorithm 1: Local Semantic Sensing

---

**Input:** All data samples  $\{x_i\}_{i=1}^N$ .

**Output:** Semantically aware local metrics  $\{M_i\}_{i=1}^N$ .

- Use unsupervised knowledge with some supervision to learn local manifold structure in the form of neighborhood graph  $W$ .
  - Determine local distances using supervision and the learned local manifold structure in  $W$ .
- 

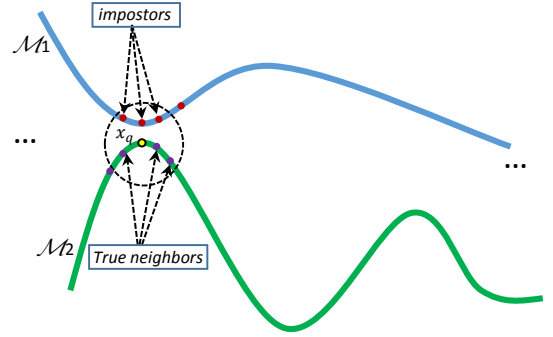


Fig. 2. An example of neighborhood impostors for a given data  $x_q$ .

depending upon which point is the query. This is based on the assumption in the paper, that distances are inherently contextual (local) in a retrieval application. It should be pointed out, that the required asymmetry in distance function computation is application-specific.

This also can be interpreted as different latent projections are applied to all data samples to measure the similarity between the query and candidates based on the intrinsic characteristic of the individual query point. In a clinic application point of view, such a projection used to retrieve similar patient record with cancer is likely to be very different from search query with AIDS. However, in applications, where the distance function is required to be symmetric, the average of the two asymmetric values can be used. The standard Euclidean or Mahalanobis distance can be thought of as extreme special cases of the point-based distance, where the same distortion matrix is used globally over the data. The matrix  $M_i$  encodes the key parameters, which need to be learned by our approach.

### C. Overview of Approach

In this section, we provide an overview of the approach used for distance function learning. The overall approach essentially contains two steps. The first step leverages the unsupervised information in order to learn the local manifold structure and encode it in the form of a graph structure  $W$ , though some supervision is used even at this stage. The second phase uses this semantic structure in conjunction with the available supervision, in order to propagate the similarity structure over the entire data set. A brief description of the approach is provided in Algorithm 1. In the following, each of the different steps of the algorithm will be described in detail.

### D. Learning Semantic Distance on Data Manifold

Most semi-supervised learning algorithms rely on good affinity matrix design, based on so-called graph preserving

criteria [29]. Similarity measures should not only utilize the affinities revealed by the user supervision, but also consider the semantic knowledge inherent in the data distribution along lower dimensional manifolds. The latter is learned most effectively from the unsupervised information, which is generally available more freely. For greater generality, we incorporate a multi-manifold model. The natural assumption here is that the semantically related data are located contiguously along a particular manifold, and the structure of the manifold has a direct impact on distance function computation. Of course, in some cases, the data points in different manifolds may be close to one another based on *traditional definitions*, but may not be semantically very related. Such data points are referred to as *impostors*. It is important to design an approach, which can use the unsupervised and supervised data in order to learn the conceptual relationships more effectively.

In order to illustrate this point, we will use a toy example. In figure 2, there are two smoothed curves indicating two 1- $d$  manifold structures embedded in 2- $d$  euclidean space. The manifold structure reflects the underlying data distributions. We assume samples from the first class lie on  $\mathcal{M}1$  (the blue curve) and the second class lie on  $\mathcal{M}2$  (the green curve). The yellow point on  $\mathcal{M}2$  represents candidate instance  $x_q$ , and the dashed circle is the contour of the Euclidian distance. It is evident that three red points from  $\mathcal{M}1$  and three purple data from  $\mathcal{M}2$  lie within these distances, though the distance is quite different in reality. The red samples are in fact impostors, and the distances should be modeled in such a way that these impostors are eliminated or de-emphasized automatically.

The key contribution of unsupervised samples is that they define the underlying data distributions. Since we will use an embedding approach which constructs a neighborhood graph, it is important to consider only sample points which lie in the same manifold for each data point  $x_i$ . Using the underlying data distribution, we are able to build a neighborhood graph only considering sample points that lie in the same manifold for each data  $x_i$ . A particularly useful result in this context is that of Elhamifar et al [8], who proposed a sparse manifold clustering and embedding (SMCE) algorithm to fit each data point to a sparse affine subspace. Let  $\bar{c}_i$  be a  $k$ -dimensional vector specific to data point  $x_i$ , with one entry for each of the  $k$  neighbors of  $x_i$ , which are denoted by  $\mathcal{N}_i$ . If the  $j$ th entry  $c_{ij}$  of  $\bar{c}_i$  is non-zero, then this  $j$ th data point lies on the same manifold as  $x_i$ . The *SMCE Assumption* is as follows:

**SMCE Assumption** For each data sample  $x_i$  drawn from manifold  $\mathcal{M}_l$  of dimensionality  $d_l$ , consider a ball of radius  $\epsilon$  centered at  $x_i$ . Then, for all  $x_i$ , there exists  $\epsilon > 0$ , such that nonzero entries of the sparsest solution for  $\bar{c}_i$  of

$$\left\| \sum_{j \in \mathcal{N}_i} c_{ij}(x_i - x_j) \right\|_2 \leq \epsilon \text{ and } \sum_{j \in \mathcal{N}_i} c_{ij} = 1 \quad (5)$$

corresponds to the  $d_l + 1$  neighbors of  $x_i$ , which also lie on the manifold  $\mathcal{M}_l$ .

More importantly, it has been shown in [8], that the solution for  $\bar{c}_i$  can be used to determine a local affine subspace for  $x_i$ , which corresponds to its manifold with the use of these  $d_l + 1$  neighbors lying on the manifold. Therefore  $\bar{c}_i$  can be viewed as a selection vector. However, since the *SMCE* algorithm only

considers a global Euclidean distance for every data point, it cannot fully leverage local distances in the embedding process. Furthermore, it cannot leverage supervised information at all. The *LSS* approach discussed in this paper can address these issues effectively, though it uses the framework provided by the *SMCE* approach as a starting point.

The first step in good local subspace determination is to select the appropriate representatives out of the  $k$ -nearest neighbors. In another words, we start the selection from a relatively large pool and then eliminate these unqualified ones. The heuristic intuition here is that by selecting a set or data points  $S$  randomly distributed around the point  $x_i$  on the same manifold, the sum of the vectors  $(x_r - x_i)/\|x_r - x_i\|$  over all  $x_r$  in the same manifold is likely to result in a vector with smaller modulus, than by picking impostor points from other manifolds. This is because it is typically possible to pick points on the manifold such that their angles are distributed along the different directions. Therefore, we set up an optimization for the selection vector  $\bar{c}_i$  on this basis. For a data sample  $x_i$  in a  $d_l$ -dimensional manifold  $\mathcal{M}_l$ , we construct a normalized basis  $U_i$ . The first step in this local analysis is to determine the  $k$  nearest neighbors ( $k \ll n$ ) using an appropriate metric. We use indices  $i_1, \dots, i_k$  to denote the  $k$ -nearest neighbors of data  $x_i$ . In order to remove the distance variations, we further normalize the difference vectors between  $x_i$  and its neighborhood set as follows:

$$U_i \triangleq \left[ \frac{x_{i_1} - x_i}{\|x_{i_1} - x_i\|}, \dots, \frac{x_{i_k} - x_i}{\|x_{i_k} - x_i\|} \right], \quad (6)$$

where  $U_i$  is a  $d \times k$  matrix, which correspond to the normalized neighbor of  $x_i$ . The purpose of the normalization is to ensure that the determination of nonzero entries of  $\bar{c}_i$  in the SMCE-computation are not dependent on absolute distances of data points from  $x_i$ . Based on the angle-based intuition discussed above, this optimization problem is formulated as follows:

$$\begin{aligned} \min_{\bar{c}_i} \quad & \|x_i - (x_i + U_i \bar{c}_i)\|_2^2 + \lambda \|\bar{c}_i\|_1 = \|U_i \bar{c}_i\|_2^2 + \lambda \|\bar{c}_i\|_1 \\ \text{s.t.} \quad & \mathbf{1}^T \bar{c}_i = 1. \end{aligned} \quad (7)$$

Once, the selection vector  $\bar{c}_i$  has been determined, the local manifold  $\mathbb{A}_i$  can be derived directly as follows:

$$\mathbb{A}_i = \{x_i + U_i \bar{c}_i \mid \bar{\mathbf{1}}^T \bar{c}_i = 1, \bar{c}_i \in \mathbb{R}^k\}. \quad (8)$$

Here the unit vector of length  $k$  is denoted by  $\bar{\mathbf{1}}$ . Another explanation behind the optimization objective in equation (7) is that the distance between  $x_i$  to its local affine subspace is minimized.

In order to further refine the process, we can add a distance-based penalty term to the afore-mentioned objective function in addition to sparse constraints on  $\bar{c}_i$ . The intuition of this term is to penalize the reconstruction of the manifold from the points far away from  $x_i$ . Furthermore, this penalty term also allows the addition of supervision at least to a minor degree, even at the stage of manifold determination, though the supervision information will be used more extensively at a later stage. This is encoded with the use of a diagonal matrix  $Q^i \in \mathbb{R}^{k \times k}$  specific to data point  $x_i$ . The  $j$ th diagonal entry of this matrix  $Q_{jj}^i$  encodes penalty information for the data point  $x_j$  in relation to  $x_i$ . This is achieved by about the distance of

data point  $x_i$  to  $x_j$ . The value of this penalty entry is defined as follows;

$$Q_{jj}^i = \frac{\|x_j - x_i\|^2}{\sum_{l=1}^k \|x_l - x_i\|^2}, \forall l = 1 \dots k \quad (9)$$

Thus, the diagonal entries of  $Q^i$  indicates the relative distance from  $x_i$  to  $x_j \in \mathcal{N}_i$ . The points closer to  $x_i$  obtain smaller values, and further points are assigned larger values. We would like to penalize reconstruction from points far away from  $x_i$ , and therefore the penalty term in the augmented objective function will be constructed, so that larger values on the diagonal entry result in larger penalties for the corresponding data point. Furthermore, in order to incorporate supervision directly at the manifold learning stage, we set  $Q_{jj}^i$  to zero or an arbitrarily small number if a positive constraint exists from  $x_i$  to  $x_j$ . On the other hand, for negative constraints, we do not wish to reconstruct the given data instance  $x_i$  from points with different semantics. This can be achieved by setting the entry in  $Q^i$  to an arbitrarily large value. Thus, the augmented objective function can be defined as follows:

$$\begin{aligned} \min_{c_i} \quad & \|U_i \bar{c}_i\|_2^2 + \lambda \sum_j Q_{jj}^i c_{ij} \\ \text{s.t.} \quad & \bar{\mathbf{1}}^T \bar{c}_i = 1, \end{aligned} \quad (10)$$

Here,  $\lambda$  is a balancing parameter to control between distances of neighborhood reconstruction and affine reconstruction error. Equation 10 is now in a standard constrained quadratic programming form, and it can be solved efficiently [33].

Once the coefficient vectors  $\{\bar{c}_i\}_{i=1}^n$  have been obtained, the normalized weight  $W$  between data points can be computed as follows:

$$w_{ij} = \left| \frac{c_{ij}/d^2(x_i, x_j)}{\sum_{t=1}^m c_{it}/d^2(x_i, x_t)} \right|. \quad (11)$$

The learned adjacency matrix  $W$  is asymmetric, with a few nonzero entries corresponding to the neighborhood relationship between  $x_i$  and others in the same manifold. The values measure the semantic affinities between data instances. This weight  $W$  encodes the relationships between data points based on local manifold behavior and (a limited amount of) supervision. In the next section, we will discuss how to work with this intermediate distance computation, and leverage supervision more effectively in order to create a more refined distance function.

### E. Leveraging Supervision

In this section, we will discuss how to use the semantic relationships  $W$  between the data points, which were learned largely with unsupervised information. This information will be combined with supervision to learn the (parameterized) local distortion matrices  $M_i$ , which were introduced in Definition 1. The key is to learn parameters which minimize the empirical loss using the supervised knowledge provided. If only the supervised information were to be used, then the similarity and dissimilarity constraints can be used in order to measure the empirical loss as follows:

$$E = f \left( \frac{1}{|S|} \sum_{(i,j) \in S} \|x_i - x_j\|_{M_i}^2 - \frac{1}{|D|} \sum_{(i,j) \in D} \|x_i - x_j\|_{M_i}^2 \right), \quad (12)$$

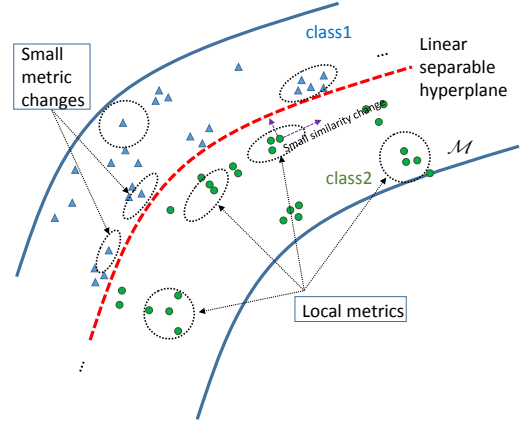


Fig. 3. A geometric intuition of the use of supervised information and local metric propagation.

where  $f$  can be any nonnegative increasing function to ensure the minimization function is bounded below. Typical choices of  $f$  includes exponential function, logistic function and hinge function. We explicitly use the exponential function throughout this paper. However, this approach does not leverage the semantic similarities encoded in the matrix  $W$  effectively, and it only learns the local functions at the data points, where the supervised similarity constraints are specified. The goal here is to use the learned semantic relationships, in conjunction with smoothness constraints. In other words, the distortion matrices of two semantically similar point which are adjacent to one another on the manifold should be similar. This can be partially achieved by adapting an approach discussed in [31] as follows:

$$R = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n W_{ij} \|M_i - M_j\|_F^2, \quad (13)$$

This approach propagates the supervised knowledge to the unsupervised instances. However, such an approach is still insufficient for learning good metrics at unsupervised data instances. The reason is that small changes in  $M_i$  has no direct relation to learning distance due to the isotropic penalization of the Frobenius norm. A simple example is that  $M_j$  could change arbitrarily refer to  $M_i$  with a fixed norm difference. Therefore, we need to add more constraints to ensure distances will change more consistently with the asymmetric semantic relationships  $W_{ij}$  implied by the manifolds. In account for this, we set up the following term  $P$  as

$$P = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|W_{ij} d_{M_i}^2(x_i, x_j) - W_{ji} d_{M_j}^2(x_i, x_j)\|_2^2. \quad (14)$$

A geometric explanation is illustrated in Figure 3, where two dimensional linearly separable data is embedded along smooth manifolds in a three dimensional space. The red dashed line indicates the separating plane for binary classes (blue triangles are from the first class, and green circles are from the second class). The black dashed circles demonstrate local metrics for some arbitrary data samples. The points close to the separating plane tend to have ellipsoidal metrics. It because, the similarity changes rapidly in the perpendicular direction of the separating plane. In contrast, similarity changes more slowly in the perpendicular direction along the separating plane. The points far away from the separation boundary exhibit an almost circular (euclidian) distortion metric due to

isotropy of the local affinities. The different components discussed above can be integrated together to create a combined objective function:

$$\min_{M_1, \dots, M_n} E(M_i) + \gamma R(M_i) + \mu P(M_i) + \frac{\nu}{2} \sum_{i=1}^n \|M_i\|_F^2 \quad (15)$$

*s.t.*  $M_1, \dots, M_n \geq 0,$

Here,  $\gamma$  and  $\mu$  are balancing coefficients. The notation  $\|M_i\|_F$  equals to  $\sqrt{\text{tr}(M_i^T M_i)}$  that represents the Frobenius matrix norm, and  $\text{tr}(M_i)$  denotes the trace of matrix  $M_i$ . Since the data lies on multiple low dimensional smoothed manifolds, we can work with the lower dimensional reduction in the learning process. Therefore, we parameterize distance  $d(x_i, x_j)$  using a projection matrix  $L_i$  according to Eq.(2) and the exponential function as  $f$  in Eq.(12). Then the objective becomes,

$$\begin{aligned} \min_{L_1, \dots, L_n} & \exp\left(\frac{1}{2|\mathcal{S}|} \sum_{(i,j) \in \mathcal{S}} (x_i - x_j)^T L_i L_i^T (x_i - x_j)\right) \\ & - \frac{1}{2|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} (x_i - x_j)^T L_i L_i^T (x_i - x_j) \\ & + \frac{\gamma}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|W_{ij} d_{L_i}^2(x_i, x_j) - W_{ji} d_{L_j}^2(x_i, x_j)\|_2^2 \\ & + \frac{\mu}{n^2} \sum_{i=1}^n \sum_{j=1}^n W_{ij} \|L_i - L_j\|_F^2 + \frac{\nu}{2} \sum_{i=1}^n \|L_i\|_F^2. \end{aligned} \quad (16)$$

The optimization objective is bounded below. The gradient of each term is:

$$\begin{aligned} \frac{\partial E(L)}{\partial L_i} &= E(L) \left( \frac{1}{|\mathcal{S}|} \sum_j H_{ij} L_i \mathbb{1}\{x_i \in \mathcal{S}\} - \frac{1}{|\mathcal{D}|} \sum_j H_{ij} L_i \mathbb{1}\{x_i \in \mathcal{D}\} \right), \\ \frac{\partial R(L)}{\partial L_i} &= \frac{2}{n^2} \sum_{j=1}^n (L_i - L_j) W_{ij} - (L_j - L_i) W_{ji}, \\ \frac{\partial P(L)}{\partial L_i} &= \frac{4}{n^2} \sum_{j=1}^n (C_{ij} W_{ij} H_{ij} - C_{ji} W_{ij} H_{ij}) L_i, \\ \frac{\partial F(L)}{\partial L_i} &= L_i, \end{aligned} \quad (17)$$

Here  $H_{ij} \in \mathbb{R}^{d \times d}$  is a rank one matrix defined by  $(x_i - x_j)(x_i - x_j)^T$ . The notation  $C_{ij}$  represents a weighted difference  $d_{M_i}(x_i, x_j)$  and  $d_{M_j}(x_i, x_j)$ , and is equal to  $W_{ij}(x_i - x_j)^T L_i L_i^T (x_i - x_j) - W_{ji}(x_i - x_j)^T L_j L_j^T (x_i - x_j)$ . The aforementioned method can be easily solved via coordinate descent method [18].

### III. DISCUSSION

In this section, we discuss two important issues for the proposed LSS model. The first is the intrinsic dimension determination for each manifold. We explicit incorporate multi-manifold assumption for sample with same semantic lies on a same smooth manifold. Determining the exact dimension of each manifold can be useful for learning metric with low rank representation. It is not only useful in term of efficient computation but also can make LSS robust to ambient noise. The second is out-of-sample handling. Most existing manifold learning approaches have no direct extension for efficient unseen sample handling. In contrast, our proposed approach supports dynamic updates, rather than learning from scratch.

#### A. Discussion of Intrinsic Dimension for LSS

The advantage of utilizing SMCE assumption is that it provides information about intrinsic dimension of the manifolds. This information can be further leveraged to learn proper low rank metrics for each data points. It comes from the assumption that for a given sample  $x_i \in \mathcal{M}_i$  and its neighbors in  $\mathcal{M}_i$  approximatively lie in a  $d_i$  dimensional tangent space of  $\mathcal{M}_i$  at  $x_i$ . Moreover, the non-zero coefficient of affine space construction coefficient  $c_i$  is expected to have  $d_i + 1$  nonzero entries, which correspond to  $d_i + 1$  linear dependent in its tangent space. The method suggested by [8] for finding the intrinsic manifold dimension by first applying spectral clustering method on  $W$ , and take the median of number of non-zero coefficients of each sample with a data cluster. The detailed description is out of the scope of this paper. We refer interested readers to [8]. An advantage of this approach is that the low dimensional projection  $L_i$  can be systematically suggested by the intrinsic dimension of each manifold for different semantics. Our LSS method could suggest different dimension for data from different manifolds to calculate the local projection matrix  $L_i$ . On the other hand, a global method would be inflexible, because of its inability to adapt to the differential intrinsic dimensionality of different localities.

#### B. Discussion of Out-of-Sample LSS

Many existing manifold learning and locality sensitive techniques do not naturally contain a out-of-sample extension. For in-sample data classification and retrieval, we can construct the neighborhood graph and learn LSS modal via the proposed intension propagation framework. However, when a new data appears, it is not desirable to build all models from scratch. The extension for out-of-sample can be efficiently performed via Eq. (10) to construct an incrementally updated neighborhood graph with respect to the new involved samples only while keep the learned coefficient unchanged. Then the metrics for the unseen sample can be learned by Eq. (16), while keeping other metrics fixed. The metric of the newly added point can be efficiently learned using existing knowledge from the newly added points.

## IV. EXPERIMENTAL RESULTS

In this section, we present experiments by comparing the proposed LSS approach with other state-of-the-art algorithms on various data sets. Our general goal will be to pick a diversity of both baseline methods, and data sets in order to show how the different scenarios affect the various baselines in a differential way. We demonstrate the advantages of our approach, and also show that our approach is generally more robust to different scenarios because of its ability to combine the supervised and unsupervised data in a locally optimized way.

#### A. Data Sets

In order to demonstrate the robustness and effectiveness of the proposed LSS method in context of retrieval, we conduct our experiments from four different data sources. These different data sources are quite different, and tend to behave quite differently for different algorithms. This is important from the perspective of illustrating the robustness of our approach. The

specific sources were the *UCI machine learning repository* [1], *ORL face recognition data set* [23], *USPS hand written digit recognition data set* [13], as well as *COREL image retrieval data set* [6].

- 1) **UCI Benchmark Data Set** [1]: We use *Iris*, *Wine* and *Letter* as the benchmark to compare the proposed method with other off-the-shelf algorithms. Irish contains 150 four dimensional samples in three classes. Wine is similar to Irish contains 178 samples with three classes in a 13 dimensional space. The last, letter provides 20,000 instances with 16 attributes. And there are 26 different classes assigned to these data points.
- 2) **ORL Face Database** [23]: The ORL dataset is a commonly used face recognition data set. It contains a set of face images taken between April 1992 and April 1994 at AT&T Laboratories Cambridge. The dataset contains ten different images of each of 40 distinct subjects with time, lighting, facial expression, and facial detail variations. All the images use a dark homogenous background with a tolerance of rotation of face up to 20 degree. We follow the pre-processing procedure shown in [5] cropping image to  $32 \times 32$  pixels, with 256 gray levels per pixel. Each image is represented as a 1024-dimensional vector. We further reduce the dimensionality to 50 by using PCA. The data set is available at [5].
- 3) **USPS Handwritten Digit Dataset** [13]: We use the most popular subset from USPS database. It contains 9,298  $16 \times 16$  handwritten digit images. Similar to face images, handwritten digit images have been widely used for manifold ranking task [32]. We randomly selected 200 samples per digit to test our purposed algorithm.
- 4) **COREL Image Retrieval Dataset** [6]: The image data set consists of 2,000 images published by COREL Corporation. Images have resolution of  $384 \times 256$  or  $256 \times 384$ . The *COREL* image database is widely used in many data mining tasks including, CBIR [16], manifold ranking [28], distance learning [19], multiple-instance learning [6]. We follow [34] to extract high level image features based on SIFT descriptors. The dimensionality was reduced to 50 to accelerate learning. Sample images are illustrated in Figure 4.

## B. Baselines

We compared our method extensively with many state-of-the-art algorithms including:

- **Euclidean Metric**: We denote this metric by the acronym *EU*.
- **Mahalanobis**: We denote this metric by the acronym *Mah*. Here,  $M$  is the inverse covariance matrix from all data samples.
- **ITML**[7]: This Information Theoretic Metric Learning algorithm learns the Mahalanobis distance by minimizing differential relative entropies under pairwise relevance constraints.



Fig. 4. Sample images from COREL image database, randomly sampled from concept “sunset scenes”, “battle ships” and “beach”.

- **DCA**[12]: Discriminative Component Analysis improves RCA[24] by exploring negative constraints.
- **SSMetric**[11]: Semi-supervised Distance metric learning, which incorporate both knowledge from neighborhood graph and sparse labeled constraints.
- **ISD**[35]: Instance specific semi-supervised distances metric learning. It propagates metrics via a smoothness regularization term.
- **EMR**[28]: Efficient manifold ranking, which is of the most recent work in manifold ranking.

## C. Evaluation Metrics

A wide variety of measures such as precision, recall, F-measure, mean average precision (MAP) and normalized discounted cumulative gain (NDCG) [17] are available. However, not all the aforementioned measures are meaningful in context of content based retrieval and web search. In general, for any kind of information or multimedia retrieval task, the retrieval engine typically presents the top  $k$  results. In this context, the top- $k$  precision effectively measures the retrieval performance. The value of  $k$  is set within the range of 20 to 100. The precision for a particular value of  $k$  is defined as follows:

$$P_k = \frac{|\text{Relevant Documents}|}{k}.$$

We average over all different query samples, and report the mean precision as the evaluation metric. The variation over different values of  $k$  can also be used.

## D. Results

In this section, we report our results compared to other state-of-the-art methods. All the results are averaged over 10 different runs.

1) *UCI Benchmark Data Sets*: For each data set, we randomly selected 5% of the data from each class set as supervised nodes. In order to mimic real feedback from a retrieval system, we selected a fixed number of constraints

TABLE I. RETRIEVAL PERFORMANCE FOR EACH UCI DATA SET. THE BEST PERFORMANCE IS IN BOLD.

UCI Performance Method / Rank	Irish			Wine			Letter		
	Rank 1	Rank 10	Rank 20	Rank 1	Rank 10	Rank 20	Rank 1	Rank 10	Rank 20
EU	95.92	94.15	93.13	84.83	71.85	65.96	92.62	78.86	69.40
Mah	95.24	91.77	88.50	88.20	74.04	66.10	91.41	78.45	69.97
ITML	96.60	94.35	92.21	85.96	70.00	64.83	94.50	84.34	77.54
DCA	95.24	94.56	93.16	86.52	80.51	77.53	91.12	81.38	74.94
SSMetric	95.92	94.97	93.98	89.33	75.62	70.20	92.54	78.24	69.54
ISD	97.28	94.42	93.10	83.17	73.88	69.72	93.45	80.14	70.86
EMR	97.28	93.40	91.50	85.39	70.73	64.55	86.05	79.15	73.30
LSS	<b>97.96</b>	<b>95.13</b>	<b>94.05</b>	<b>93.26</b>	<b>84.94</b>	<b>82.05</b>	<b>95.24</b>	<b>86.22</b>	<b>80.63</b>

per data record in order to provide pairwise constraints. For the *Irish* and *Wine* data sets, the sample size is relatively small, therefore, only 5 constraints were selected for each node. Furthermore, we use  $k = 20$ , which is typical for a real retrieval system. It is evident from Table I, the proposed *LSS* method outperforms the baselines significantly in all three benchmarks. It is worth mentioning that, some of the compared methods performed even worse than the trivial Euclidean metric because of the over-fitting caused by limited labeled data. This shows that it is sometimes very tricky to use purely supervised methods without the benefit of unlabeled data in many scenarios.

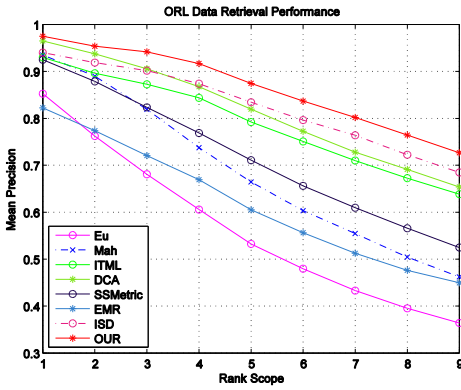


Fig. 5. ORL face image retrieval result.

2) *ORL Face Data Set*: We implemented our proposed method on the *ORL* and reports precision various as the scope of rank changes in figure 5. We provide only a sample from each class with supervised information within the 10 nearest neighbors. Since the *ORL* dataset only contains 10 samples per class, we only report the precision curve using the top 9 points, excluding the query point itself. Compared to other global metric learning algorithms, our method and *ISD* perform the best. The *EMR* method performs the worst because of its reliance on large amounts of supervision, which are not available in this case.

3) *USPS Digit*: The *USPS* hand-written digit recognition data set is also used for algorithmic comparison. We consider all digits from 0 to 9, and randomly selected 200 samples from each class. We used 5% samples from each class used as supervised node, and provide 10 samples per node as the user-specified constraints. Unlike the precision curve obtained from the *ORL* face data set, *EMR* performs relatively well with other methods, while *ISD* drops its precision significantly. It is worth mentioning, that the unsupervised Mahalanobis performs the worst, because the second order covariance cannot correctly characterize the locally distorted distance direction in the

absence of supervision. We are able to show only a part of the curve in this case, because the performance has already dropped out of our region of interest. This underlines the point made earlier about the *robustness* of our approach to a variety of data sets, because of its ability to combine unsupervised data, supervised data and locality in an optimal way, which does not seem to be well addressed by the other methods.

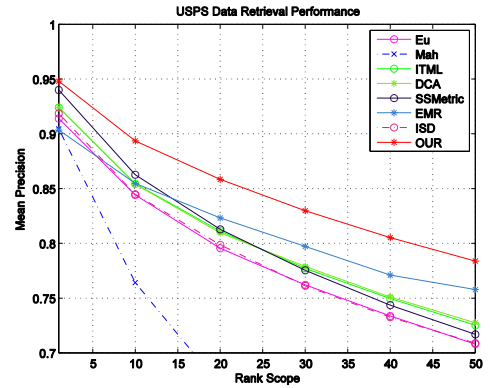


Fig. 6. USPS handwritten digit image retrieval result.

4) *COREL Natural Images*: Finally, we illustrated the proposed method on the *COREL* image retrieval data set. We used 5% data and 10 constraints for each of these items, in order to construct supervised user intentions. Since this data set is somewhat larger, we used the top 50 retrieval results. In figure 7 we report the top 50 retrieval results for each of the methods. We obtain similar results as the *USPS* data set for the standard Mahalanobis approach, which performs quite poorly. The performance of the other methods are very close to each other. They are either slightly better or worse than the Euclidean metric. However, our proposed algorithm receives higher improvement than any other method, and we outperformed the second best method by more than 5% at all retrieval levels.

### E. Parameter Sensitivity

Model selection plays a key role in many learning models, and the performance of an algorithm may therefore vary with parameter changes. In this subsection, we evaluate the performance of the proposed *LSS* methods with two main parameter changes. To demonstrate model selection, we use a 500 sample *USPS* subset, which each class contain 50 samples. All other parameters were fixed, and we varied  $\gamma$  from 100 to  $10^7$ . The corresponding rank 1 precision change is illustrated in figure 8. It is evident that rank 1 precision increases with  $\gamma$ . After  $\gamma$  increases beyond 100, the *LSS* approach performs extremely well. Similarly, Figure 9 illustrates the effect of  $\mu$  on the

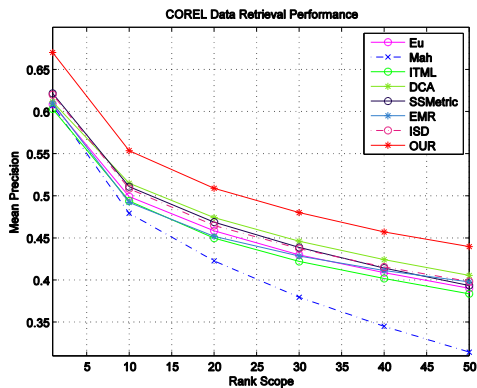


Fig. 7. COREL image retrieval result.

final result. Typically, we obtain the best performance within the range of 100 to 50,000. Across all of our experiments, we selected our parameters  $\gamma$  and  $\mu$  from the aforementioned range by using a small subset of the training records for model selection methodology, based on retrieval performance.

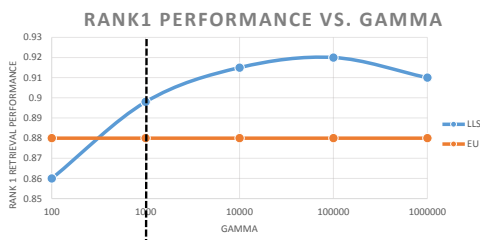


Fig. 8. Rank 1 performance vs. gamma changes.

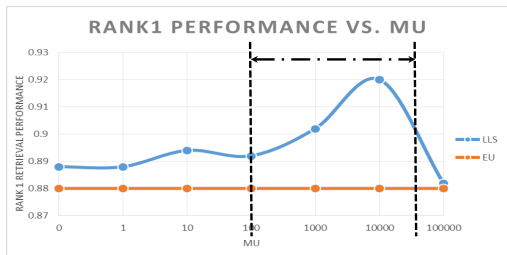


Fig. 9. Rank 1 performance vs. mu changes.

## V. RELATED WORK

In recent year, emerging research interests have been focused on learning a good distance metric in feature space and it has been shown distance metric is crucial in many real-world applications. A fundamental problem in many data mining and web search applications is that we are confronted to high dimensional input data so that Euclidean distance is not as meaningful as it in a low dimensional space due to the so-called curse of dimensionality. Many literature are trying to find a low-dimensional representations given high-dimensional input data. In this section, we will summarize some important work in both metric learning and manifold learning. A detailed survey are available at [30] for interested reader.

### A. Metric Learning

Existing metric learning [2] algorithms can be splitting into different groups with different criteria: (1) linearity: linear and nonlinear; (2) supervision: supervised, semi-supervised, and unsupervised; (3) locality: global and local. The most existing method fall in the region of supervised global linear metric learning, which learns a PSD matrix or a low-dimensional projection matrix follows the definition shown in Eq. (2). It try to learn metrics by keeping similar points close while pushing dissimilar samples further. Linear Discriminant Analysis (LDA) [9] is one of the most popular supervised linear embedding method, which seeks for the projection direction such that same classes are clustered and different classes are well separated. Comparing the global method of LDA, LMNN [26] is a local supervised metric learning approach, which aims to find projection direction where the local class discriminability is maximized. However, both LDA and LMNN are using explicit class label information, which is rarely available in web search applications. In contrast, Relevant Component Analysis (RCA) [24], DCA [12], Neighborhood Component Analysis (NCA)[10], and ITML [7] only using pairwise constraints or called side-information for learning a proper distance metric. Nevertheless, these methods are still insufficient for dealing with only limited supervised samples. SSMetric [11] and  $S^3$ Metric [15] are graph preserving based approaches which propagate label information through an affinity graph to alleviate the amount of supervision acquired. But, these two models are lack of considering the nonlinearity of the data distribution, which a single metric hardly to handle (toy example shown in figure 1). Many cases, multiple metrics or even instance based metric are desired, one prominent method that follows this line of thinking is ISD [35]. ISD learns a instance specific metric learning, and instead of propagating label information, it provides each data instance a local metric. The unsupervised data point acquire supervision via a smoothness term, which assume metrics will not change significantly in a local region. However, ISD did incorporate manifold learning to a metric adaptation framework, or even consider local semantic ambiguities. In this work, our method is a local, nonlinear, and semi-supervised approach learning semantic sensitive metrics.

### B. Manifold Learning

Manifold learning can be also seen as a kind of unsupervised metric learning. The main idea is to learn an underlying low-dimensional manifold where geometric relationships between most of the observed data are preserved. Manifold learning usually coherent with unsupervised metric learning and dimensionality reduction. It is essentially to learn a mapping function, such that distance between two points in the original space can be represented by standard Euclidean metric in the new space. To do so, the first step of most dimensionality reduction methods is to build a neighborhood graph. Local methods, such as LLE [22], and Laplacian Eigenmaps [3] try to preserve local relationships. Global methods, such as Isomap3 [4] and Semidefinite Embedding [27] try to preserve local and global relationships among all data points. SMCE [8] utilize multiple nonlinear assumption cluster data and reduce dimensionality simultaneously. However, SMCE is unsupervised method, which is not specifically designed for metric learning purpose. We adopt SMCE assumptions,

and seamlessly integrate neighborhood constructors with supervised user intension to construct local concept sensitive metrics.

## VI. CONCLUSIONS

In this paper, we proposed a novel LSS approach within semi-supervised metric learning framework. We explicitly consider semantic and locality information from both supervised and unsupervised data in order to create a robust and locality sensitive metric. This method harnesses the latent semantic knowledge in the unsupervised data in conjunction with limited intensional information. From a conceptual perspective, each instance “senses” its local neighborhood semantic with user-specified supervision in order to construct a semantically aware distance function. We conduct experiments on a wide range of data sets, and show the robustness and effectiveness of our approach over a large number of state-of-the-art baseline methods.

## ACKNOWLEDGMENT

This work was founded in part to Shiyu Chang and Thomas S. Huang by the National Science Foundation under Grand No. 1318971 and the Samsung Global Research Program 2013 under Theme “Big Data and Network”, Subject “Privacy and Trust Management In Big Data Analysis”. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

- [1] D. N. A. Asuncion. UCI machine learning repository, 2007.
- [2] C. C. Aggarwal. Towards systematic design of distance functions for data mining applications. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 9–18. ACM, 2003.
- [3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, 2001.
- [4] M. Bernstein, V. D. Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds, 2000.
- [5] D. Cai, X. He, Y. Hu, J. Han, and T. S. Huang. Learning a spatially smooth subspace for face recognition. In *CVPR*, 2007.
- [6] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, Dec. 2006.
- [7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, Corvallis, Oregon, USA, 2007.
- [8] E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 55–63. 2011.
- [9] K. Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.
- [10] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2004.
- [11] S. C. H. Hoi, W. Liu, and S.-F. Chang. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *TOMCCAP*, 6(3), 2010.
- [12] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, CVPR '06*, pages 2072–2078, Washington, DC, USA, 2006. IEEE Computer Society.
- [13] J. J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(5):550–554, May 1994.
- [14] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3610–3617. IEEE, 2013.
- [15] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu. Semi-supervised sparse metric learning using alternating linearization optimization. In *ACM KDD Conference*, pages 1139–1148, New York, NY, USA, 2010. ACM.
- [16] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1):262–282, Jan. 2007.
- [17] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [18] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
- [19] G.-J. Qi, C. C. Aggarwal, and T. S. Huang. Transfer learning of distance metrics by cross-domain metric sampling across heterogeneous spaces. In *SDM*, pages 528–539, 2012.
- [20] B. Qian, X. Wang, F. Wang, H. Li, J. Ye, and I. Davidson. Active learning from relative queries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1614–1620. AAAI Press, 2013.
- [21] B. Qian, X. Wang, J. Wang, H. Li, N. Cao, W. Zhi, and I. Davidson. Fast pairwise query selection for large-scale active learning to rank. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 607–616. IEEE, 2013.
- [22] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.
- [23] F. S. Samaria, F. S. S. \*t, A. Harter, and O. A. Site. Parameterisation of a stochastic model for human face identification, 1994.
- [24] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis, 2002.
- [25] F. Wang, J. Sun, and S. Ebadollahi. Composite distance metric integration by leveraging multiple experts’ inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining*, 5(1):54–69, 2012.
- [26] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *In NIPS*. MIT Press, 2006.
- [27] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Comput. Vision*, 70(1):77–90, Oct. 2006.
- [28] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, and J. Luo. Efficient manifold ranking for image retrieval. In *ACM SIGIR Conference*, 2011.
- [29] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):40–51, Jan. 2007.
- [30] L. Yang. Distance metric learning: A comprehensive survey, 2006.
- [31] D. C. Zhan, M. Li, Y. F. Li, and Z. H. Zhou. Learning instance specific distances using metric propagation. In *ICML Conference*, pages 1225–1232, New York, NY, USA, 2009. ACM.
- [32] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- [33] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, and J. Ye. Feafiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1034–1042. ACM, 2013.
- [34] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang. Hierarchical gaussianization for image classification. In *ICCV*, 2009.
- [35] Z.-H. Zhou and H.-B. Dai. Query-sensitive similarity measure for content-based image retrieval. In *ICDM Conference*, pages 1211–1215, 2006.