

# Hierarchical Large Margin Nearest Neighbor Classification

Qiaona Chen, Shiliang Sun

Department of Computer Science and Technology

East China Normal University

500 Dongchuan Road, Shanghai 200241, China

Email: slsun@cs.ecnu.edu.cn

**Abstract**—Distance metric learning has exhibited its great power to enhance performance in metric related pattern recognition tasks. The recent large margin nearest neighbor classification (LMNN) improves the performance of  $k$ -nearest neighbor classification by learning a global distance metric. However, it does not consider the locality of data distributions, which is crucial in determining a proper metric. In this paper, we propose a novel local distance metric learning method called hierarchical LMNN (HLMNN) which first builds a hierarchical structure by grouping data points according to the overlapping ratios defined by us and then learns distance metrics sequentially. Experimental results on real-world data sets including comparisons with the traditional  $k$ -nearest neighbor and the state-of-the-art LMNN show the effectiveness of the proposed HLMNN.

**Keywords**—distance metric learning; global metric; hierarchical structure;  $k$ -nearest neighbor; local metric

## I. INTRODUCTION

Distance metric learning aims to improve the performance of metric related pattern recognition algorithms by learning an appropriate metric from data. As some standard metrics (e.g., the Euclidean metric) treat all features equally and fail to reveal the structure information embedded in data, it is well motivated to develop data-driven metrics by the techniques of distance metric learning [1], [2], [3]. In this paper, we focus on supervised distance metric learning.

Supervised distance metric learning can be subdivided into two categories:

- Global distance metric learning. The basic idea is to minimize the distance between examples in the same classes and maximize the distance between examples in different classes in a global sense.

Some global distance metric learning methods [4], [5], [6] learn a single linear transformation in a subspace to minimize a cost function. These algorithms usually depend on gradient descent methods, and thus are prone to local minima.

- Local distance metric learning. Local structures in data are exploited for discriminant analysis by this kind of approaches.

For example, a local flexible distance metric using SVMs was proposed in [7]. The relationship between the  $i$ th dimension and the posterior class probabilities of class  $j$  was explored by chi-square distance in [8].

Hastie and Tibshirani [9] learned a distance metric to modify the neighborhood. Yang et al. [10] employed eigenvector analysis and bound optimization techniques to learn a local distance metric.

The recent method of large margin nearest neighbor classification (LMNN) [11] has shown its effectiveness in improving the accuracy of  $k$ -nearest neighbor classification ( $k$ NN) by the large margin principle to enlarge the between-class distance [12]. Since it uses semidefinite programming (SDP) [13] to solve a convex problem, LMNN avoids the problem of local minima and can find the optimal linear transformation.

However, LMNN is a global distance metric learning method. It can not satisfactorily represent the local metrics that are optimal in different regions of the input space. Motivated by this consideration, here we propose a new local distance metric learning method called hierarchical LMNN (HLMNN). We define a measurement called the overlapping ratio to build a hierarchy through which distance metrics are learnt sequentially.

The remainder of this paper is organized as follows. Section II briefly reviews LMNN, the foundation of the proposed HLMNN. Section III introduces HLMNN in detail. Experimental results are reported in Section IV after which we conclude the paper in Section V.

## II. OVERVIEW OF LMNN

Suppose we have a set of inputs  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  ( $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$ ) with labels  $y_i$  ( $i = 1, 2, \dots, n$ ). The goal of LMNN is to learn a global linear transformation  $L$  used in the following transformed distance

$$D(\mathbf{x}_i, \mathbf{x}_j) = \|L(\mathbf{x}_i - \mathbf{x}_j)\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^T L^T L (\mathbf{x}_i - \mathbf{x}_j),$$

where  $L$  is a  $d \times d$  matrix.

For each input  $\mathbf{x}_i$ , LMNN specifies  $k$  target neighbors which are  $k$  other inputs with the same label as  $\mathbf{x}_i$ . A binary variable  $\eta_{ij} \in \{1, 0\}$  is used to indicate whether or not input  $\mathbf{x}_i$  is a target neighbor of input  $\mathbf{x}_j$ . The objective function of LMNN has two competing terms [11]. The first term penalizes large distances between each input and its target neighbors. The second term, corresponding to a margin condition similar to that of SVMs, penalizes small distances

between each input and all other inputs that do not share the same label. Specifically, the objective function is

$$\varepsilon(L) = \sum_{ij} \eta_{ij} \|L(\mathbf{x}_i - \mathbf{x}_j)\|^2 + c \sum_{ijl} \eta_{ij}(1 - y_{il}) [1 + \|L(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \|L(\mathbf{x}_i - \mathbf{x}_l)\|^2]_+, \quad (1)$$

where positive constant  $c$  controls the relative importance of these two competing terms, binary variable  $y_{il} \in \{1, 0\}$  indicates whether or not input  $\mathbf{x}_i$  and input  $\mathbf{x}_l$  belong to the same class, and function  $[z]_+ = \max(z, 0)$  denotes the standard hinge loss.

LMNN adopts SDP to optimize the distance metric learning criterion under a convex problem formulation. The resulting SDP problem is

$$\begin{aligned} \min & \sum_{ij} \eta_{ij} (\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j) + \\ & c \sum_{ijl} \eta_{ij}(1 - y_{il}) \varepsilon_{ijl} \\ \text{s.t.} & \begin{cases} (\mathbf{x}_i - \mathbf{x}_l)^T M (\mathbf{x}_i - \mathbf{x}_l) - (\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j) \\ \geq 1 - \varepsilon_{ijl}, \\ \varepsilon_{ijl} \geq 0, \\ M \geq 0, \end{cases} \end{aligned} \quad (2)$$

where matrix  $M = L^T L$ , constraint  $M \geq 0$  requires matrix  $M$  to be positive semidefinite, and  $\varepsilon_{ijl}$ 's are slack variables [11].

### III. HLMNN

As data distributions in the real world often exhibit localities, global distance metric learning methods including LMNN can not satisfactorily represent the local distance metrics which are respectively optimal in different regions of the input space.

Consider an example shown in Figure 1 where the distribution of six classes exhibits a high locality. Clearly, these six classes can be grouped into two clusters. Regions with high localities may correspond to local optimal metrics which are very different from the global optimal metric. To validate this, we perform distance metric learning as in LMNN for each of the two clusters, and obtain two optimal metrics as

$$\begin{pmatrix} 0.332 & -0.007 \\ -0.007 & 0.339 \end{pmatrix} \quad (3)$$

and

$$\begin{pmatrix} 0.488 & 0.040 \\ 0.047 & 0.118 \end{pmatrix}. \quad (4)$$

Obviously, these two metrics are very different. It means that these two clusters contain different types of local discriminant information. A global metric learned in LMNN can not satisfactorily represent these two local metrics.

We now proceed to present the details of the proposed HLMNN. In short, HLMNN has four steps. Firstly, overlapping ratios between every two classes are calculated.

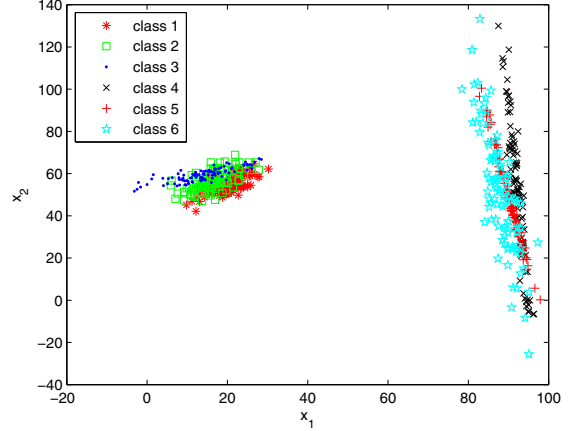


Figure 1. An example of data distribution with six classes where localities are observed.

Secondly, a hierarchical structure is built through grouping classes according to their overlapping ratios. Thirdly, hierarchical metrics are learned by LMNN as a subroutine. Finally, test points are classified in a hierarchical manner. Below we detail the key elements of these four steps.

#### A. Overlapping ratio

Here we define a new data clustering criterion—overlapping ratio which is not sensitive to outliers and has no requirement for the shape of data distributions.

For each input  $\mathbf{x}_i$ ,  $k$  target neighbors are specified as in Section II. If the distance between  $\mathbf{x}_i$  and differently labeled  $\mathbf{x}_j$  does not exceed the distance between  $\mathbf{x}_i$  and its  $k$ th nearest target neighbor,  $\mathbf{x}_j$  is defined to be the active neighbor of  $\mathbf{x}_i$ . Suppose  $active_{ij}$  denotes the number of points in class  $j$ , which can be regarded as the active neighbors of some input in class  $i$ . It is calculated as:

$$active_{ij} = \sum_{t_1=1}^n \sum_{t_2 \neq t_1}^n [\sigma(y_{t_1} - i) \times \sigma(y_{t_2} - j) \times h(L\|\mathbf{x}_{t_1} - \hat{N}_{t_1}\|^2 - L\|\mathbf{x}_{t_1} - \mathbf{x}_{t_2}\|^2)] \quad (5)$$

where variable  $\hat{N}_{t_1}$  represents the  $k$ th nearest target neighbor of  $\mathbf{x}_{t_1}$ ,

$$\sigma(r) = \begin{cases} 1, & \text{for } r = 0 \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

and

$$h(r) = \begin{cases} 1, & \text{for } r > 0 \\ 0, & \text{for } r \leq 0. \end{cases} \quad (7)$$

The active neighbor number reflects the possibility of the inputs in class  $i$  are likely to be wrongly labeled as some points in class  $j$ . Suppose  $num_i$  denotes the number of points in class  $i$ . The overlapping ratio between class  $i$  and

class  $j$  is defined as:

$$OLR_{ij} = \frac{active_{ij} \times active_{ji}}{num_i \times num_j}. \quad (8)$$

In this paper, overlapping ratios are used to measure the overlapping degrees between classes and detect the regions exhibiting high localities. If the overlapping ratio between some classes is large, the local metric for these classes tends to be very different from the global metric for all classes and in this case learning a global metric is not optimal.

### B. Clustering

Hierarchical clustering is used to build clusters of classes according to their overlapping ratios. Before this clustering, overlapping ratios between every two classes are computed and the nonzero overlapping ratios are sorted in descending order. Then, this hierarchical clustering starts with each class as a cluster. Firstly, classes represented by a high overlapping ratio are found. Then, these classes are merged into the same cluster if they are not in the same cluster. These two steps are repeated until the overlapping ratio considered is less than a given threshold.

The major steps of clustering for a general problem are shown by the follow pseudocode, where  $z$  is a constant for stopping clustering and vector  $\{a_1, a_2, \dots, a_p\}$  contains all nonzero overlapping ratios between classes in descending order.

- **begin initialize**  $k \leftarrow 0$ ,  $cluster_i \leftarrow class_i$  ( $i = 1, 2, \dots, m$ )
  - **do**  $k \leftarrow k + 1$
  - find the nearest classes according to the overlapping ratio  $a_k$ , say  $class_i$  and  $class_j$
  - find the clusters which  $class_i$  and  $class_j$  belongs to, say  $cluster_i$  and  $cluster_j$
  - merge  $cluster_i$  and  $cluster_j$ ,  $m \leftarrow m - 1$
  - **until**  $a_k < z$  or  $m = 1$
- **return** the clusters
- **end**

As described above, the procedure of clustering terminates when the overlapping ratio  $a_k$  is smaller than the specified constant  $z$  or the cluster number  $m$  is 1. It is obvious that the choice of  $z$  is very important. If  $z$  is too large, each class will form a cluster in which case hierarchical distance metric learning is pointless. If  $z$  is too small, all classes are merged into one cluster and HLMNN degenerates to LMNN.

### C. Hierarchical Distance Metric Learning

After clustering, a hierarchical structure has been built. As the name implies, HLMNN learns metrics in a hierarchical way, and it considers local metrics as well as the global metric. The resultant hierarchical distance metrics are composed of a global metric and its sub-metrics (local metrics). The global metric is learned by taking each cluster as a class and its sub-metrics are learned for every cluster. Metrics here are learned by the original LMNN algorithm.

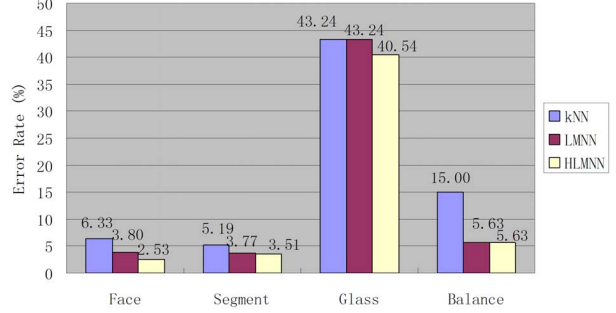


Figure 2. The test error rates of  $k$ NN, LMNN and HLMNN on different data sets.

Table I  
PROPERTIES OF DATA SETS

	Face	Segment	Glass	Balance
Size (train)	320	1540	140	465
Size (test)	79	770	74	160
Classes	40	7	6	3
Features	50	19	10	4

### D. Classification

The test data points are also classified hierarchically. Firstly, inputs are mapped into a feature space by the global metric. In this space, the test points are classified into a cluster by  $k$ NN. Then, the inputs are re-mapped into another feature space by the sub-metric, where the test data points are further classified by  $k$ NN.

## IV. EXPERIMENT

Our experiment employs four real-world data sets from different domains. The face data set is taken from Cambridge university computer laboratory. The segment, glass, and balance data sets are taken from UCI machine learning repository. Table I summarizes the main properties of the data sets.

We compared the proposed method against two established approaches. The first baseline approach is  $k$ NN based on the Euclidean distance. The second baseline is the state-of-the-art method—LMNN. The nearest neighbor number and the target neighbor number are both taken to be 3 in our experiments as in [11].

The classification errors are shown in Figure 2. For the first three data sets, HLMNN outperforms the other two methods. For the data set balance, HLMNN has the same performance with LMNN, which can be explained by

Table II  
THE OVERLAPPING RATIOS FOR THE BALANCE DATA SET

	Class 1	Class 2	Class 3
Class 1	-	13.10	12.54
Class 2	-	-	0.15
Class 3	-	-	-

exploring the overlapping ratios for different classes in this data set.

Table II shows that class one has large overlapping ratios with both class two and class three. In this case, all the three classes are merged into the same cluster, and therefore HLMNN degenerates to LMNN and they lead to the same performance.

#### V. CONCLUSION

This paper has proposed a novel hierarchical distance metric learning method—HLMNN. We have defined a new criterion—overlapping ratio for grouping classes in HLMNN. Experimental results on different data sets show that the HLMNN method outperforms the traditional  $k$ NN and recent LMNN methods. Theoretical analysis of HLMNN and experiments on more data sets will be our future work.

#### ACKNOWLEDGMENT

The authors would like to thank the National Natural Science Foundation of China and Shanghai Educational Development Foundation for funding respectively under Project 60703005 and 2007CG30.

#### REFERENCES

- [1] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Technical Report*, Michigan State University, 2006.
- [2] S. Sun and Q. Chen. Kernel regression with a mahalanobis metric for short-term traffic flow forecasting. *Lecture Notes in Computer Science*, 5326:9–16, 2008.
- [3] J. Tenenbaum, V. Silva and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [4] J. Goldberger, S. Roweis, G. Hinton and R. Salakhutdinov. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, 17:513–520, 2004.
- [5] A. Hillel, T. Hertz, N. Shental and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.
- [6] A. Holub, Y. Liu and P. Perona. On constructing facial similarity maps. *Proceedings of Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [7] C. Domeniconi, D. Gunopulos and J. Peng. Large margin nearest neighbor classifiers. *IEEE Transactions on Neural Networks*, 16(4):899–900, 2005.
- [8] C. Domeniconi, J. Peng and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1281–1285, 2002.
- [9] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, 1996.
- [10] L. Yang, R. Jin, R. Sukthankar and Y. Liu. An efficient algorithm for local distance metric learning. *Proceedings of the National Conference on Artificial Intelligence*, pp. 543–548, 2006.
- [11] K. Weinberger, J. Blitzer and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, 18:1475–1482, 2006.
- [12] S. Sun. Local within-class accuracies for weighting individual outputs in multiple classifier systems. *Pattern Recognition Letters*, 31(2):119–124, 2010.
- [13] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.