© International Journal of Combinatorial Optimization Problems and Informatics, Vol. 5, No. 1, Jan-April 2014, pp. 15-20. ISSN: 2007-1558.

Learning a Kernel Matrix Using Some Similar and Dissimilar pairs

Hamideh Hajiabadi

Birjand University of Technology

hajiabadi@birjandut.ac.ir

Abstract

A lot of machine learning algorithms are based on metric functions, which good functions lead to better results. Distance metric learning has been widely attracted by researchers in last decade. Kernel matrix is somehow a distance function which indicates the similarity between two instances in the feature space which contains high dimensions. Traditional distance metric learning approaches are based on Mahanalobis distance which result in optimizing a positive semi definite problem. This kind of approaches need high computational time and do not work well in the case of data with high dimensions. Another filed which is involved by researchers in last decade is building a good kernel matrix which separate non separable data best. This paper proposed a new algorithm in order to learn kernel matrix which is based on distance metric learning. It is implemented and applied to several standard data sets and the results are shown.

Keywords: Distance metric learning, Kernel Learning, Kernel function, Mahanalobis distance

1. Introduction

Metric functions play key role in a wide machine learning application such as classification, clustering, object recognition, data retrieval, etc. Hence; it has been particularly attracted recently. The performance of some traditional clustering or classification algorithms such as k-NN or k-means extremely depends upon a metric which calculates the distance between samples. In the last few years researcher interested in DML algorithm optimizing cost function under several type of constraints. [13] However recent results demonstrate that the proper distance metric can be learned using some side information identifying some similar or dissimilar pairs. Traditional works concentrate on Mahanalobis distance represented by:

$$d(x,y) = \|x - y\|_A = (x - y)^T A(x - y)$$

Specifying some similar and dissimilar pairs a semi definite optimization problem was defined and the best matrix A is achieved [11][12][17]. Large Margin Nearest Neighbor (LMNN) [6] [7] is an approach relied on Mahanalobis distance trying to make instances in the same class closer and those in different classes separate by a margin. Xing [8] proposed a method based on Mahanalobis distance. This method minimizes the sum square of similar pairs subjected to keep sum square of dissimilar pairs larger than predefined threshold. Schultz organized a method [5] to learn a distance metric from relative comparison such as "A is closer to B than A is closer to C" by solving a quadratic optimization problem. Davis proposed [1] an information-theoretic approach to learn a Mahalanobis distance Metric. The problem has been expressed as the minimization of the relative differential entropy between two multivariate Gaussians constrained on the distance function. Bar-Hillel [2] expressed Relevant Component Analysis (RCA) algorithm in order to learn a Mahalanobis metric based on information theory. Liu Yang [9] presented a Bayesian framework for distance metric learning by estimating a posterior distribution from the pairwise labels. [10] In [14] generalization error of DML formulation which is not depends on data dimensions is minimized and the metric is investigated.

Some other researches concentrate on risk minimization, the metric is learnt by Empirical Risk Minimization (ERM) framework. [15][16]
Solving the semi positive definite problem and the high computation complexity is the weakness of the earlier works; hence, the distance metric learning cannot be applied to data with high dimensions.

Received Nov 20, 2013 / Accepted Dec 19, 2013

The other field that attracted researcher is Kernel function. Generally, kernel functions are used in order to separate the non-separable data. Some works are performed on Distance Metric Learning in feature space. Jain [3] produced a novel approach to learn a semi definite matrix from side information in feature space. He as a first pioneer discovered the relation between distance learning and Kernel Learning via linear transformation by minimizing the LogDet divergence, subject to some constrains. [18]

Researchers are interested in finding the kernel that separates the data best. One possible way is to apply several kernels for which the best one is dominated by using cross validation technique, which is very time consuming. The other way which some researchers followed, is composite kernel. Based on given side information, Yan [4] proposed an approach to combine multiple kernels. In this paper a kernel function such as RBF is used, and the kernel matrix (gram matrix) is generated. The coefficients are learned using distance metric learning approach.

In section II distance metric learning approach are briefly described. Section III contains the proposed work in detail. Experimental result are illustrated in section IV and conclusion and future work are enclosed in section V.

2. Distance Metric Learning

Initially, the metric is briefly explained. In mathematics function $d: R^k \times R^k \rightarrow R$. d is metric where for all $x, y, z \in R^k$ following conditions are satisfied:

1- $d(x,y) \geq 0$
2- $d(x,y) = 0 \Rightarrow x = y$
3- $d(x,y) = d(y,x)$
4- $d(x,z) \leq d(x,y) + d(y,z)$

Euclidean distance is one of the function satisfying above conditions.

$$d_{Euc} = \|x - y\|_2$$

The distance also satisfies the conditions for which is defined as:

$$d_{Maha} = \|x - y\|_{\Sigma} = (x - y)^T M (x - y)$$

If M be the inverse of covariance matrix, the distance is named Mahalanobis, But M can also be a semi definite matrix which satisfies the:

$$M = L^T L$$

The Mahalanobis distance satisfies conditions 1, 3 and 4 but not 2; hence it is named as pseudometric. If M be identity matrix the distance will be Euclidean distance.

In distance metric learning given n instance $X = \{x_1, x_2, \ldots, x_n\}$ where $x_i \in R^d$, aimed to learn the semi definite matrix $M \succcurlyeq 0$ subject to reserving sum of dissimilarity pairwise distances. The instances in the same class are considered as similar and in the different classes as dissimilar. Considering two bounds $l < u$ the pairs x_i, x_j are found as similar if $d(x_i, x_j) < u$ and they are considered as dissimilar if $d(x_i, x_j) > l$.

3. The Proposed algorithm

In the proposed algorithm the similar and dissimilar pairs are identified in the original space.

$S = \{(x_i, x_j) | x_i \in X \text{ and } x_j \in X \text{ known as similar pair}\}$

$D = \{(x_i, x_j) | x_i \in X \text{ and } x_j \in X \text{ known as dissimilar pair}\}$

Next, using a kernel function, data are mapped to a Hilbert space which is named also the feature space. The kernel function which is used should satisfy several conditions known as Mercer's conditions.

$$K_{ij} = \phi(x_i)^T \phi(x_j)$$

Matrix K is symmetric semi positive definite. K can be decomposed as:

$$Kv_i = \lambda_i v_i$$

$$K[v_1, v_2, \dots, v_d] = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{bmatrix} [v_1, v_2, \dots, v_d]$$

$$\Rightarrow K = \sum_{i=1}^{d} \lambda_i v_i v_i^T$$

Let $R_i = v_i v_i^T$ which is a matrix with $n \times n$ dimensions. K can be written as $K = \sum_{i=1}^{d} \lambda_i R_i$. It means K is equal to sum of weighted R_i.

Since K is semi positive definite then $\lambda_i \geq 0$ for all $1 \leq i \leq d$. The Euclidean distance between x and y in the feature space can be written as:

$$d(x_i, x_j) = \left\| \phi(x_i) - \phi(x_j) \right\|_2^2 = K_{ii}^2 - 2K_{ij}^2 + K_{jj}^2 = \sum_{i=1}^{n} \lambda_i \left((R_k)_{ii} - 2(R_k)_{ij} + (R_k)_{jj} \right)$$

Distance metric learning approaches are aimed to learn a metric to make similar pairs closer and dissimilar pairs further apart. The goal function can be written like:

$$J = \min \sum_{(x_i, x_j) \in S} d(x_i, x_j) - \sum_{(x_i, x_j) \in D} d(x_i, x_j)$$

Sum of dissimilar pairs are added to the subject function with negative coefficient, which causes sum of similar pairs to be minimized and dissimilar pairs maximized.

K can be regarded as $K = \sum_{i=1}^{d} \alpha_i R_i, \alpha_i \geq 0$ and the problem is to find α_i in order to minimize the cost function.

$$P = \min \sum_{k=1}^{d} \alpha_k \sum_{(x_i, x_j) \in S} \left((R_k)_{ii} - 2(R_k)_{ij} + (R_k)_{jj} \right) - \sum_{k=1}^{d} \alpha_k \sum_{(x_i, x_j) \in D} \left((R_k)_{ii} - 2(R_k)_{ij} + (R_k)_{jj} \right)$$

After minimizing the subject function, α_i would be zero. Hence the constraint $\sum_{i=1}^{d} \alpha_i = c$ is added, where c is a constant given by the user.

$$S_k = \sum_{(i,j) \in S} \left((R_k)_{ii} - 2(R_k)_{ij} + (R_k)_{jj} \right)$$

$$D_k = \sum_{(i,j) \in D} \left((R_k)_{ii} - 2(R_k)_{ij} + (R_k)_{jj} \right)$$

The cost function can be written as:

$$P = \min \quad \sum_{k=1}^{d} \alpha_k (S_k - D_k)$$

$$s.t. \sum_{i=1}^{d} \alpha_k = c$$

$$, \alpha_k \geq 0$$

By giving

$$A = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{bmatrix}, X_k = S_k - D_k, X = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}$$

The cost function can be written as:

$$LP = \min X^T \times A$$

$$s.t. A \times 1 = c,$$

$$A > 0$$

The LP is a kind of linear programming and can be solved by traditional Simplex.

4. Experimental results

The proposed metric learning approach explained in previous section in details. The constrained optimization problem lead to a linear optimization; hence, the Simlpex can solve it. Consequently the computational time of the proposed algorithm is extremely low in comparison with the traditional ones. It can be done by O(n).

The data is divided into two groups: training data and test one. The training data is used to extract the kernel matrix, then in order to classify test data, the 1NN algorithm is applied. The data set used for this reason is Iris data set. The iris database are demonstrated in figure 1.

Table 1 data set's propoerties

Data set	Samples	Dimensions	Number of classes
Iris	150	4	3

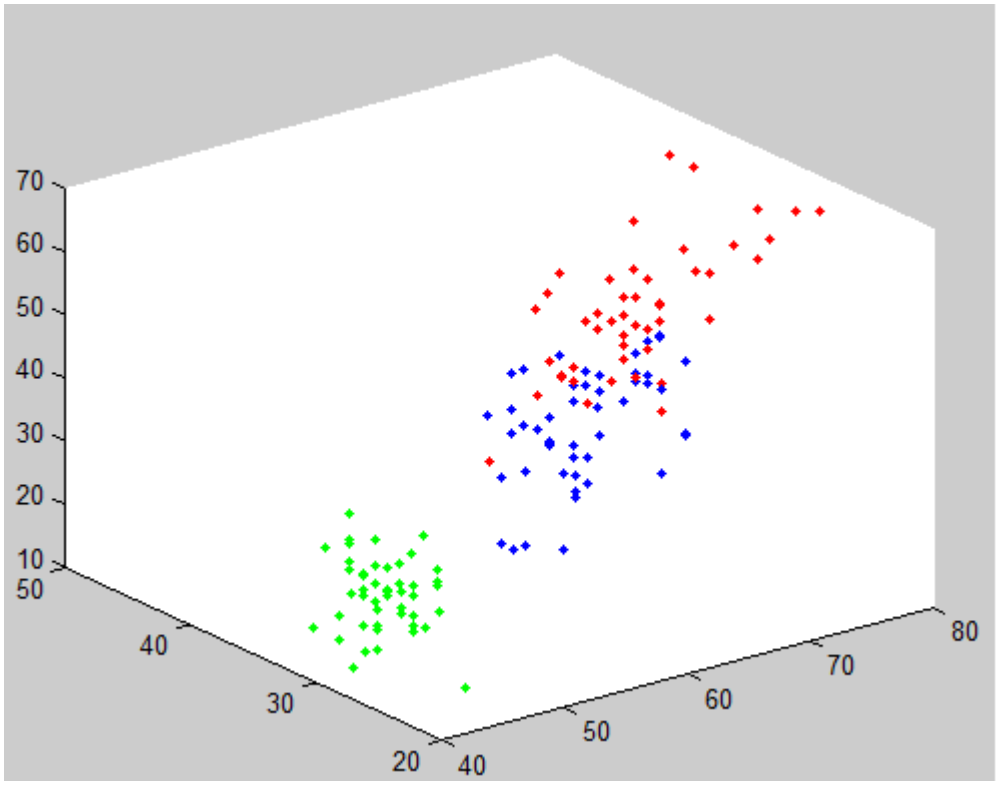

Figure 1 Iris data set presented separately per class.

Data divided into third partition, one of them are considered as training and the remaining part as test data. First the 1NN algorithm are applied to original data, the number of misclassifies are shown in table 1. Then the 1NN classify is applied to the kernel matrix attained by proposed distance metric algorithm. The results are shown in table 2 expressing the proposed algorithm decreases the number of misclassifies.

Table 2 the results obtained by proposed algorithm and are compared with original one.

	Original data	Proposed Metric learning
Number of misclassify in 1NN	4	2
Number of misclassify in 5NN	3	0
Number of misclassify in kMeans clustering	16	5

REFERENCES

1. [Dav07] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "*Information theoretic metric learning*," in Proc. 24th Int. Conf. Mach. Learn., pp. 209–216, **2007.**
2. [Hil05] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "*Learning a Mahalanobis metric from equivalence constraints*," J.Mach. Learn. Res.,
vol. 6, pp. 937–965, Dec. **2005**.
3. [Jai12] P. Jain, B. Kulis, J. V. Davis, I. S. Dhillon; "*Metric and Kernel Learning Using a Linear Transformation". J. Mach. Learn. Res.*, vol. 13, pp. 519−547, Mar. **2012**
4. [Mik10] F. Yan, K. Mikolajczyk, J. Kittler, and M. A. Tahir, "*Combining Multiple Kernels by Augmenting the Kernel Matrix*," in Proc. Multiple Classifier Systems, pp. 175-184, **2010**.
5. [Sch04] M. Schultz and T. Joachims, "*Learning a distance metric from relative comparisons*," in Proc. Adv. Neural Inf. Process. Syst., pp. 1–8, **2004**.

6. [Wei05] K. Q. Weinberger, J. Blitzer, and L. K. Saul (2006). In Y. Weiss, B. Schoelkopf, and J. Platt (eds.), "*Distance Metric Learning for Large Margin Nearest Neighbor Classification***",** Advances in Neural Information Processing Systems 18 (NIPS-05). MIT Press: Cambridge, MA, **2005**.

7. [Wei09] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "*Distance metric learning for large margin nearest neighbor classification*," J. Mach. Learn. Res., vol. 10, pp. 207–244, Feb. **2009**

8. [Xin02] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "*Distance metric learning, with application to clustering with side-information*," in Proc. Adv. Neural Inf. Process. Syst., pp. 505–512, **2002**.

9. [Yan07] L. Yang, R. Sukthankar, and R. Jin, "*Bayesian active distance metric learning*," in Proc. 23rd Conf. Uncertainty Artif. Intell., Vancouver, BC, Canada, **2007**, pp. 1–8.

10. [Yan10] F. Yan, K. Mikolajczyk, J. Kittler, and M. A. Tahir, "*Combining Multiple Kernels by Augmenting the Kernel Matrix*," in Proc. Multiple Classifier Systems, pp. 175-184, **2010**.

11. A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. 20th Int. Conf. Mach. Learn.*, **2003**, pp. 11–18.
12. A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *J. Mach. Learn. Res.*, vol. 6, pp. 937–965, Dec. **2005**.
13. L. Yang and A. R. Jin, "Distance metric learning: A comprehensive survey," Dept. Comput. Sci. Eng., Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep., **2006**.
14. R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning: Theory and algorithm," in *Advances in Neural Information Processing*. Fort Atkinson, WI, USA: Curran Associates Inc., **2009**, pp. 862–870.
15. W. Bian and D. Tao, "Learning a distance metric by empirical loss minimization," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, vol. 2. **2011**, pp. 1186–1191.
16. W. Bian and D. Tao, "Constrained empirical risk minimization framework for distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1194–1205, Aug. **2012**.
17. B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2143–2157, Dec. **2009**.
18. V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, nos. 5–6, pp. 544–557, **2009**.