Extending the Relevant Component Analysis Algorithm for Metric Learning Using Both Positive and Negative Equivalence Constraints

Dit-Yan Yeung Hong Chang

Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Corresponding author: Dit-Yan Yeung, dyyeung@cs.ust.hk, +852-2358-1477 (fax)

Abstract

Relevant component analysis (RCA) is a recently proposed metric learning method for semi-supervised learning applications. It is a simple and efficient method that has been applied successfully to give impressive results. However, RCA can make use of supervisory information in the form of positive equivalence constraints only. In this paper, we propose an extension to RCA that allows both positive and negative equivalence constraints to be incorporated. Experimental results show that the extended RCA algorithm is effective.

Key words: metric learning, Mahalanobis metric, semi-supervised learning.

1 Introduction

Many pattern recognition algorithms rely on some metric or non-metric distance functions either explicitly or implicitly. Very often, the performance of an algorithm depends critically on how good the distance function is. Instead of predefining a distance function based on some prior knowledge about the application at hand, a more appealing approach is to learn an appropriate distance function automatically using supervisory information available.

Distance function learning for supervised learning applications has a long history which can be dated back to the 1970s. However, distance function learning for semi-supervised learning applications has only been studied recently. Among the algorithms proposed, *relevant component analysis* (RCA) is a simple and efficient algorithm proposed by Bar-Hillel et al. [1,2] for the learning of Mahalanobis metrics in a semi-supervised fashion using positive equivalence constraints. However, RCA cannot make use of negative equivalence constraints which may also be available in some applications.

In the next section, we briefly summarize the RCA algorithm and propose an extension to it that can incorporate both positive and negative equivalence constraints in a natural way. We then show empirically in Section 3 that the extended RCA algorithm can outperform the original algorithm in situations where the negative equivalence constraints available can provide useful supervisory information that cannot be provided by positive equivalence constraints.

2 RCA and Our Extension

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a data set of n points in some input space and $\mathcal{C}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}\}, i = 1, \ldots, k$, be k chunklets, where $\mathbf{x}_{ij}$ denotes the jth point in the ith chunklet and m_i is the number of points in the ith chunklet. Let $m = \sum_{i=1}^{k} m_i$. If each of the n points is represented by a node in an undirected graph and each positive equivalence constraint, given in the form of a point pair $(\mathbf{x}_r, \mathbf{x}_s)$, is represented by an undirected edge connecting nodes r and s, then each chunklet is simply a connected component of the graph.

The original RCA algorithm first computes the so-called within-chunklet covariance matrix as

$$\mathbf{C} = \frac{1}{m} \sum_{i=1}^{k} \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)^T, \qquad (1)$$

where $\mathbf{m}_i$ denotes the sample mean of the ith chunklet and $\mathbf{v}^T$ denotes the transpose of vector $\mathbf{v}$. Based on $\mathbf{C}$, a whitening transformation matrix $\mathbf{W}$ is computed as $\mathbf{W} = \mathbf{C}^{-\frac{1}{2}}$ and each input data point $\mathbf{x}_i$ is linearly transformed to $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i$. The subsequent pattern recognition task, such as clustering, is then performed on the transformed data set $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. In case the input dimensionality is large compared with the number of points in the chunklets, $\mathbf{C}$ is singular and hence its inverse does not exist. In this case, [2] proposed applying a dimensionality reduction step before performing the whitening transformation.

We propose in this paper an extension to RCA that allows both positive and negative equivalence constraints to be used. Let $\mathcal{S}$ and $\mathcal{D}$ denote the sets of positive and negative equivalence constraints, respectively. We define the following matrix based on $\mathcal{S}$ as:

$$\begin{aligned}\mathbf{C}_{\mathcal{S}} &= \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \left[\left(\mathbf{x}_i - \frac{\mathbf{x}_i+\mathbf{x}_j}{2}\right)\left(\mathbf{x}_i - \frac{\mathbf{x}_i+\mathbf{x}_j}{2}\right)^T + \right. \\ &\qquad\qquad \left. \left(\mathbf{x}_j - \frac{\mathbf{x}_i+\mathbf{x}_j}{2}\right)\left(\mathbf{x}_j - \frac{\mathbf{x}_i+\mathbf{x}_j}{2}\right)^T\right] \\ &= \frac{1}{2|\mathcal{S}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \end{aligned} \tag{2}$$

where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. Note that this form is similar to that above by treating each pair in $\mathcal{S}$ as a chunklet. This slight variation makes it easier to extend RCA to incorporate negative equivalence constraints as well into metric learning. Similar to $\mathbf{C}_{\mathcal{S}}$, we define the following matrix based on $\mathcal{D}$:

$$\mathbf{C}_{\mathcal{D}} = \frac{1}{2|\mathcal{D}|} \sum_{(\mathbf{x}_k,\mathbf{x}_l)\in\mathcal{D}} (\mathbf{x}_k - \mathbf{x}_l)(\mathbf{x}_k - \mathbf{x}_l)^T. \tag{3}$$

Let $\mathbf{W}_1 = \mathbf{C}_{\mathcal{S}}^{-\frac{1}{2}}$ and $\mathbf{W}_2 = \mathbf{C}_{\mathcal{D}}^{\frac{1}{2}}$. We then define the linear transformation matrix $\mathbf{W}$ as

$$\mathbf{W} = \mathbf{W}_2\mathbf{W}_1 = \mathbf{C}_{\mathcal{D}}^{\frac{1}{2}}\mathbf{C}_{\mathcal{S}}^{-\frac{1}{2}}. \tag{4}$$

Each input data point $\mathbf{x}_i$ is first linearly transformed to $\mathbf{z}_i = \mathbf{W}_1\mathbf{x}_i$ using the positive equivalence constraints. $\mathbf{z}_i$ is then linearly transformed to $\mathbf{y}_i = \mathbf{W}_2\mathbf{z}_i = \mathbf{W}\mathbf{x}_i$. This extension of the original RCA algorithm has been inspired by the classical linear feature extraction method called Fisher discriminant analysis (FDA) and its kernel extension called kernel Fisher discriminant analysis (KFD). Both attempt to maximize the Fisher criterion, which is expressed as a Rayleigh quotient, resulting in maximizing the between-class scatter and minimizing the within-class scatter simultaneously. It is interesting to note that while applying $\mathbf{W}_1$ is analogous to reducing within-class scatter, applying $\mathbf{W}_2$ is analogous to increasing between-class scatter. The illustrative examples in Section 3.1 further elaborate these properties of our extended RCA algorithm.

Similar to the original RCA algorithm, dimensionality reduction may also be applied if necessary.

3 Experiments

In this section, we perform some experiments on both toy data and real data to demonstrate the efficacy of the proposed method.

3.1 Illustrative Examples

Figure 1 shows the performance of the extended RCA algorithm on a toy data set consisting of two elongated Gaussian distributed classes as shown in Figure 1(a). The positive and negative equivalence constraints are represented as point pairs in Figure 1(b) and (d), respectively. We center each point pair by subtracting the mean from the points it contains, and represent each covariance matrix as an ellipse. Figure 1(c) and (e) show the centered pairs together with the covariance ellipses corresponding to $\mathbf{C}_{\mathcal{S}}$ and $\mathbf{C}_{\mathcal{D}}$, respectively. The data set after applying the extended RCA transformation is shown in Figure 1(f). Notice that the positive equivalence constraints do not help much because the covariance ellipse corresponding to $\mathbf{C}_{\mathcal{S}}$ is close to a circle. On the other hand, the negative equivalence constraints can play a complementary role.

Figure 2 shows another illustrative example which is similar to the toy data set in [1] where standard RCA with positive equivalence constraints only can learn a good metric. We assume that only negative equivalence constraints are available as shown in Figure 2(a). Our extended RCA algorithm can still lead to compact clusters in the transformed data set as shown in Figure 2(c).

3.2 Semi-Supervised Clustering on UCI Data

Besides the toy examples, we also assess the effectiveness of the extended RCA algorithm indirectly by how much it can improve the clustering results in semi-supervised clustering tasks with both positive and negative equivalence constraints. We use the Euclidean distance without metric learning for baseline comparison. In summary, the following five distance measures for the k-means clustering algorithm are included in our comparative study: (1) k-means without metric learning; (2) k-means with standard RCA for metric learning (only $\mathcal{S}$ provided); (3) k-means with extended RCA for metric learning (only $\mathcal{D}$ provided); (4) k-means with extended RCA for metric learning (both $\mathcal{S}$ and $\mathcal{D}$ provided); (5) same as (2) but with more positive equivalence constraints (both $\mathcal{S}$ and $\mathcal{S}'$ provided). We set $|\mathcal{S}| = |\mathcal{S}'| = |\mathcal{D}|$ for all clustering tasks. As in [1,3,4], we use Rand index as the clustering performance measure. For each data set, we randomly generate 20 different $\mathcal{S}$ ($\mathcal{S}'$) and/or $\mathcal{D}$ sets as equivalence constraints. In addition, for each $\mathcal{S}$ ($\mathcal{S}'$) and/or $\mathcal{D}$ set, we perform 20 runs of k-means with different random initializations and report the average Rand index over the 20 runs.

We perform semi-supervised clustering experiments on the following six data sets from the UCI Machine Learning Repository: Iris (150/4/3/30), Wine (178/13/3/20), Ionosphere (351/34/2/30), Boston housing (506/13/3/40), Balance (625/4/3/40), and Diabetes (768/8/2/50). The numbers inside the brackets ($n/d/c/t$) refer to the number of data points n, number of features d, number of clusters c, and number of randomly selected positive or negative equivalence constraints $t = |\mathcal{S}| = |\mathcal{S}'| = |\mathcal{D}|$.

Figure 3 summarizes the k-means clustering results based on different distance measures as numbered above. From the results, we can see that standard RCA generally improves the clustering performance. Extended RCA with only negative constraints is generally worse than standard RCA with only positive constraints. From the results of paired t-test with significance level 0.05, we can conclude that the extended RCA algorithm with both positive and negative equivalence constraints gives the best results for the Iris, Ionosphere and Diabetes data sets, while it is statistically comparable to standard RCA with the same number of equivalence constraints for the other three data sets.

4 Conclusion

In this paper, we have proposed the extended RCA algorithm that can incorporate both positive and negative equivalence constraints. The extension is natural and very effective. Experimental results on both toy and real data show that the extended RCA algorithm can outperform the original algorithm in situations where the negative equivalence constraints available can provide useful supervisory information not available from positive equivalence constraints.

Acknowledgments

The research described in this paper has been supported by Competitive Earmarked Research Grant (CERG) HKUST6174/04E from the Research Grants Council of the Hong Kong Special Administrative Region, China.

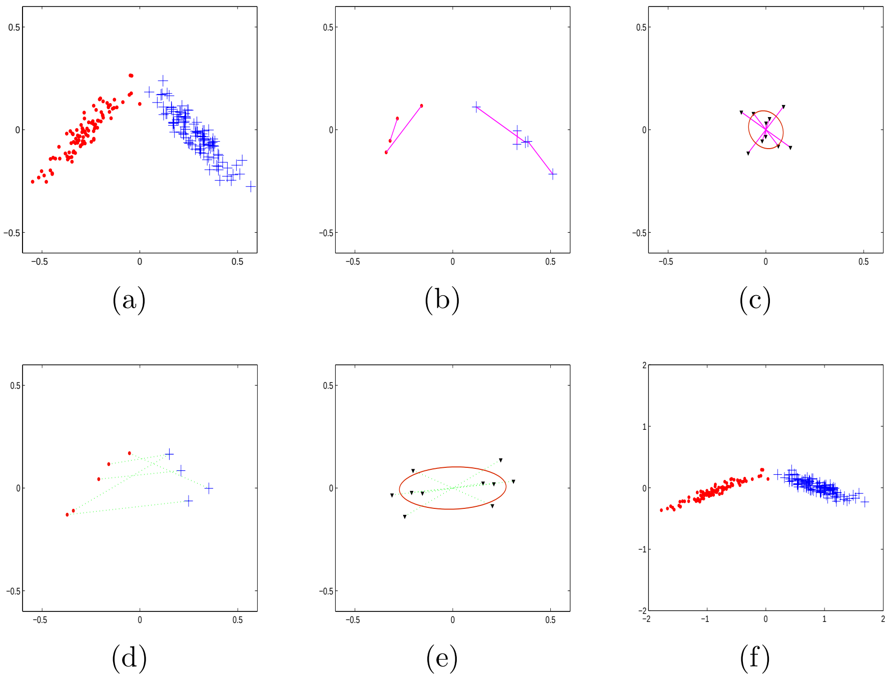
References

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 11–18, Washington, DC, USA, 21–24 August 2003.

[2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.

[3] H. Chang and D.Y. Yeung. Locally linear metric adaptation for semi-supervised clustering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 153–160, Banff, Alberta, Canada, 4–8 July 2004.

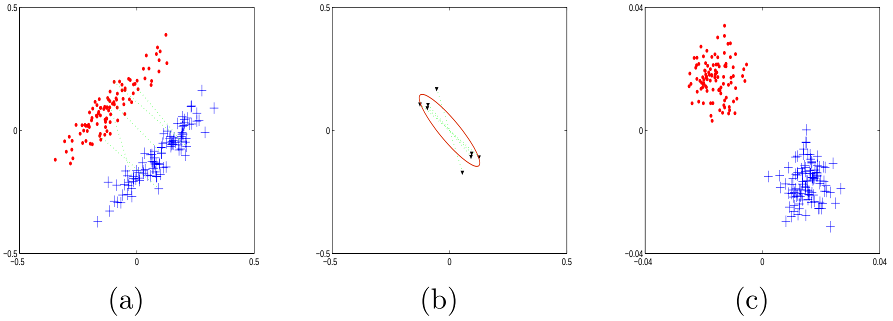
[4] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, Cambridge, MA, USA, 2003.



(a) (b) (c)

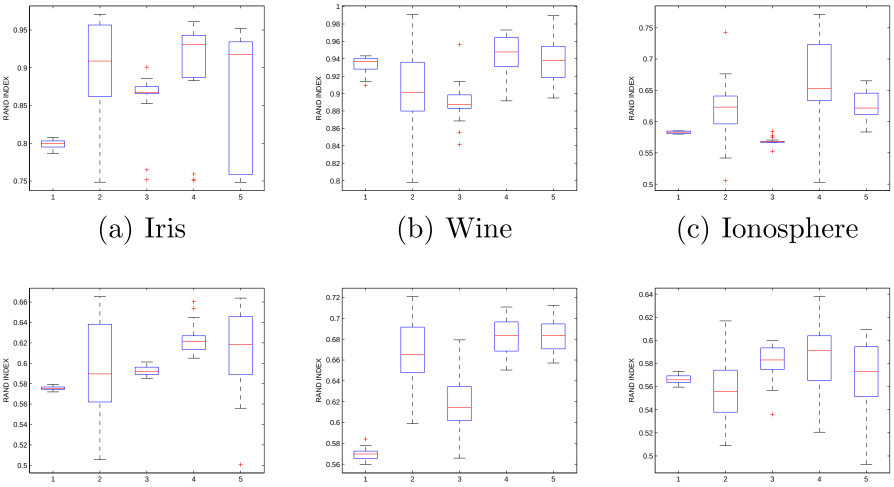
(d) (e) (f)

Fig. 1. Extended RCA algorithm applied to toy data set 1: (a) original data set with two classes; (b) point pairs for positive equivalence constraints; (c) centered pairs for (b) and the covariance ellipse; (d) point pairs for negative equivalence constraints; (e) centered pairs for (d) and the covariance ellipse; (f) data set after applying extended RCA transformation.



(a) (b) (c)

Fig. 2. Extended RCA algorithm applied to toy data set 2: (a) original data set with negative equivalence constraints shown; (b) centered pairs and the covariance ellipse; (c) data set after transformation.



(a) Iris

(b) Wine

(c) Ionosphere

(d) Boston housing

(e) Balance

(f) Diabetes

Fig. 3. Clustering results for six UCI data sets.