Semi-supervised Clustering Using Heterogeneous Dissimilarities

Manuel Martín-Merino

Universidad Pontificia de Salamanca
C/Compañía 5, 37002, Salamanca, Spain
`mmartinmac@upsa.es`

Abstract. The performance of many clustering algorithms such as k -means depends strongly on the dissimilarity considered to evaluate the sample proximities. The choice of a good dissimilarity is a difficult task because each dissimilarity reflects different features of the data. Therefore, different dissimilarities should be integrated in order to reflect more accurately which is similar for the user and the problem at hand.

In many applications, the user feedback or the a priori knowledge about the problem provide pairs of similar and dissimilar examples. This side-information may be used to learn a distance metric and to improve the clustering results. In this paper, we address the problem of learning a linear combination of dissimilarities using side information in the form of equivalence constraints. The minimization of the error function is based on a quadratic optimization algorithm. A smoothing term is included that penalizes the complexity of the family of distances and avoids overfitting.

The experimental results suggest that the method proposed outperforms a standard metric learning algorithm and improves the classical k -means clustering based on a single dissimilarity.

1 Introduction

Clustering algorithms such as k -means depend critically on the choice of a good dissimilarity [17]. A large variety of dissimilarities have been proposed in the literature [1]. However, in real applications no dissimilarity outperforms the others because each dissimilarity reflects often different features of the data [11]. So, instead of using a single dissimilarity it has been recommended in [10,11] to consider a linear combination of heterogeneous dissimilarities.

Several authors have proposed techniques to learn a linear combination of kernels (similarities) from the data [10,13,11,18]. These methods are designed for classification tasks and assume that the class labels are available for the training set. However, for certain applications such as Bioinformatics, domain experts provide only incomplete knowledge in the form of which pairs of proteins or genes are related [7]. This a priori information should be incorporated into semi-supervised clustering algorithms via equivalence constraints [5]. Thus, [17] proposed a distance metric learning algorithm that incorporates such similarity/dissimilarity information using a convex programming approach. The

experimental results show a significant improvement in clustering results. However, the algorithm is based on an iterative procedure that is computationally intensive particularly, for high dimensional applications. To avoid this problem, [5,8,15] presented more efficient algorithms to learn a Mahalanobis metric. However, these algorithms are not able to incorporate heterogeneous dissimilarities and rely on the use of the Mahalanobis distance that may not be appropriate for certain kind of applications.

Our approach considers that the integration of dissimilarities that reflect different features of the data should help to improve the clustering results. To this aim, a linear combination of heterogeneous dissimilarities is learnt considering the relation between kernels and distances [12]. A learning algorithm is proposed to estimate the optimal weights considering the similarity/dissimilarity constraints available. The method proposed is based on a convex quadratic optimization algorithm and incorporates a smoothing term that penalizes the complexity of the family of distances avoiding overfitting.

The algorithm has been evaluated considering several benchmark UCI datasets and two human complex cancer problems using the gene expression profiles. The empirical results suggest that the method proposed improves the clustering results obtained considering a single dissimilarity and a widely used metric learning algorithm.

This paper is organized as follows: Section 2 introduces the idealized metric considered in this paper, section 3 presents the algorithm proposed to learn a combination of dissimilarities from equivalence constraints. Section 4 illustrates the performance of the algorithm using several benchmark datasets. Finally, Section 5 gets conclusions and outlines future research trends.

2 Idealized Dissimilarity: Impact on Clustering Results

Let $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^d$ be the input patterns. We are given side-information in the form of pairs that are considered similar or dissimilar for the application at hand. Let $\mathcal{S}$ and $\mathcal{D}$ be the subset of pairs of patterns known to be similar/dissimilar defined as:

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \text{ is similar to } \mathbf{x}_j\} \quad (1)$$

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \text{ is dissimilar to } \mathbf{x}_j\} \quad (2)$$

Let $\{d_{ij}^l\}_{l=1}^M$ be the set of heterogeneous dissimilarities considered. Each dissimilarity can be embedded in a feature space via the empirical kernel map introduced in appendix A. Let K_{ij}^l be the kernel matrix that represents the dissimilarity matrix $(d_{ij}^l)_{i,j=1}^n$. The kernel function can be written as an inner product in feature space [14] $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ and therefore, it can be considered a similarity measure [15].

The ideal similarity (kernel) should be defined such that it becomes large for similar patterns and small for dissimilar ones. Mathematically, the ideal kernel is defined as follows:

$$k_{ij}^* = \begin{cases} \max_l \{k_{ij}^l\} & \text{If } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ \min_l \{k_{ij}^l\} & \text{If } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \end{cases} \quad (3)$$

The idealized kernel introduced in this paper is related to the one proposed by [2] for classification purposes: $k(\mathbf{x}_i, \mathbf{x}_j) = 1$ if $y_i = y_j$ and 0 otherwise, where y_i denotes the label of $\mathbf{x}_i$. However, there are two differences that are worth to mention. First, the ideal kernel proposed by [2] doesn't take into account the topology and distribution of the data, missing relevant information for the identification of groups in a semi-supervised setting. Second, this kernel can be considered an extreme case of the idealized kernel defined earlier and thus, more prone to overfitting.

Considering the relation between distances and kernels [14], the idealized distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be written in terms of kernel evaluations as:

$$d^{2*}(\mathbf{x}_i, \mathbf{x}_j) = \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 \quad (4)$$

$$= k^*(\mathbf{x}_i, \mathbf{x}_i) + k^*(\mathbf{x}_j, \mathbf{x}_j) - 2k^*(\mathbf{x}_i, \mathbf{x}_j) \quad (5)$$

The idealized dissimilarity, collects information from a set of heterogeneous measures adapting the metric to the problem at hand and improving the clustering results.

3 Learning a Combination of Dissimilarities from Equivalence Constraints

In this section, we present a learning algorithm to estimate the optimal weights of a linear combination of kernels from a set of similarity or dissimilarity constraints.

Let $\{k_{ij}^l\}_{l=1}^M$ be the set of kernels obtained from a set of heterogeneous dissimilarities via the empirical kernel map introduced in appendix A. If non-linear kernels with different parameter values are considered, we get a wider family of measures that includes non-linear transformations of the original dissimilarities. The kernel sought is defined as:

$$k_{ij} = \sum_{l=1}^M \beta_l k_{ij}^l, \quad (6)$$

where the coefficients are constrained to be $\beta_l \geq 0$. This non-negative constraint on the weights helps to interpret the results and guarantees that provided all the individual kernels are positive semi-definite the combination of kernels is convex and positive semi-definite [13].

The optimization problem in the primal may be formulated as follows:

$$\min_{\beta, \xi} \frac{1}{2} \|\beta\|^2 + \frac{C_S}{N_S} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \xi_{ij} + \frac{C_D}{N_D} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \xi_{ij} \quad (7)$$

$$\text{s. t. } \beta^T \mathbf{K}_{ij} \geq K_{ij}^* - \xi_{ij} \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \quad (8)$$

$$\beta^T \mathbf{K}_{ij} \leq K_{ij}^* + \xi_{ij} \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \quad (9)$$

$$\beta_l \geq 0 \quad \xi_{ij} \geq 0 \quad \forall l = 1, \dots, M \quad (10)$$

where the first term in equation (7) is a regularization term that penalizes the complexity of the family of distances, C_S and C_D are regularization parameters that give more relevance to the similarity or dissimilarity constraints. N_S , N_D are the number of pairs in $\mathcal{S}$ and $\mathcal{D}$, $\mathbf{K}_{ij} = [K_{ij}^1, \dots, K_{ij}^M]^T$, K_{ij}^* is the idealized kernel matrix and ξ_{ij} are the slack variables that allows for errors in the constraints.

To solve this constrained optimization problem the method of Lagrange Multipliers is used. Then, the dual problem becomes:

$$\max_{\alpha_{ij}, \gamma} \quad -\frac{1}{2} \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ (\mathbf{x}_k, \mathbf{x}_l) \in \mathcal{S}}} \alpha_{ij} \alpha_{kl} \mathbf{K}_{ij}^T \mathbf{K}_{kl} - \frac{1}{2} \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \\ (\mathbf{x}_k, \mathbf{x}_l) \in \mathcal{D}}} \alpha_{ij} \alpha_{kl} \mathbf{K}_{ij}^T \mathbf{K}_{kl} \quad (11)$$

$$+ \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}, \\ (\mathbf{x}_k, \mathbf{x}_l) \in \mathcal{D}}} \alpha_{ij} \alpha_{kl} \mathbf{K}_{ij}^T \mathbf{K}_{kl} - \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \alpha_{ij} \gamma^T \mathbf{K}_{ij} - \frac{1}{2} \gamma^T \gamma \quad (12)$$

$$+ \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \alpha_{ij} \gamma^T \mathbf{K}_{ij} + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \alpha_{ij} K_{ij}^* - \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \alpha_{ij} K_{ij}^*, \quad (13)$$

subject to:

$$0 \leq \alpha_{ij} \leq \begin{cases} \frac{C_S}{N_S} & \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ \frac{C_D}{N_D} & \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \end{cases} \quad (14)$$

$$\gamma_l \geq 0 \quad \forall l = 1, \dots, M, \quad (15)$$

where α_{ij} and γ_l are the lagrange multipliers. This is a standard quadratic optimization problem similar to the one solved by the SVM. The computational burden does not depend on the dimensionality of the space and it avoids the problem of local minima.

Once the α_{ij} and γ_l are computed, the weights β_l can be obtained considering $\partial L / \partial \beta = 0$:

$$\beta = \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \alpha_{ij} \mathbf{K}_{ij} - \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \alpha_{ij} \mathbf{K}_{ij} + \gamma. \quad (16)$$

The weights β_l can be substituted in equation (6) to get the optimal combination of heterogeneous kernels. Next, a kernel k -means clustering algorithm [3] is run. Notice that the learning algorithm proposed may be applied together with any clustering based on kernels or dissimilarities.

Several techniques are related to the one proposed here. In [17] it has been proposed an algorithm to learn a full or diagonal Mahalanobis metric from similarity

information. The optimization algorithm is based on an iterative procedure that is more costly particularly for high dimensional problems. [5] and [8,15] have proposed more efficient algorithms to learn a Mahalanobis metric from equivalence constraints. The first one (Relevant Component Analysis), can only take into account similarity constraints. Both of them, rely solely on a Mahalanobis metric that may fail to reflect appropriately the sample proximities for certain kind of applications. Hence, they are not able to integrate heterogeneous measures that convey complementary information. Finally, [16] has proposed a modification of the maximum margin clustering that is able to learn a linear combination of kernels. However, this algorithm is unsupervised and can not incorporate a priori information in a semi-supervised way. Besides, it can not be extended to other clustering algorithms based on dissimilarities or kernels as the method proposed here.

3.1 Support Vectors and KKT Complementary Conditions

The Lagrange Multipliers determine how difficult is for the linear combination of kernels to satisfy the constraints (8)-(9). Next, we can obtain a relation between the constraints satisfaction and the value of the Lagrange Multipliers.

The Karush-Kuhn-Tucker (KKT) complementary conditions in the primal are the following:

$$\alpha_{ij}(\beta^T \mathbf{K}_{ij} - K_{ij}^* + \xi_{ij}) = 0, \quad (x_i, x_j) \in \mathcal{S} \quad (17)$$

$$\alpha_{ij}(\beta^T \mathbf{K}_{ij} - K_{ij}^* - \xi_{ij}) = 0, \quad (x_i, x_j) \in \mathcal{D} \quad (18)$$

$$\eta_{ij}\xi_{ij} = 0, \quad (x_i, x_j) \in \mathcal{S}, (x_i, x_j) \in \mathcal{D} \quad (19)$$

$$\gamma_l \beta_l = 0, \quad \forall l = 1, \dots, M. \quad (20)$$

Now it can be easily shown the following proposition:

Proposition 1. *For all $(x_i, x_j) \in \mathcal{S}$*

$$\beta^T \mathbf{K}_{ij} \begin{cases} = K_{ij}^* & 0 < \alpha_{ij} < \frac{C_S}{N_S} \\ \geq K_{ij}^* & \alpha_{ij} = 0 \\ < K_{ij}^* & \alpha_{ij} = \frac{C_S}{N_S} \end{cases}$$

For all $(x_i, x_j) \in \mathcal{D}$

$$\beta^T \mathbf{K}_{ij} \begin{cases} = K_{ij}^* & 0 < \alpha_{ij} < \frac{C_D}{N_D} \\ \leq K_{ij}^* & \alpha_{ij} = 0 \\ > K_{ij}^* & \alpha_{ij} = \frac{C_D}{N_D} \end{cases}$$

When the Lagrange Multipliers are zero, the constraints are met with a margin equal or greater than zero. The corresponding pairs will not appear in the solution. When the Lagrange Multipliers are larger than zero, the constraints may be exactly met or violated. They are the support vectors as in the SVM. This will allow to solve the optimization problem more efficiently.

4 Experimental Results

The algorithm proposed has been evaluated considering a wide range of practical problems. Table 1 shows the features of the different datasets. We have chosen problems with a broad range of signal to noise ratio (Var/Samp.), varying number of samples and classes. The first three problems correspond to benchmark datasets obtained from the UCI database <http://archive.ics.uci.edu/ml/datasets/>. The last ones aim to the identification of complex human cancer samples using the gene expression profiles. They are available from bioinformatics2.pitt.edu.

Table 1. Features of the different datasets considered

	Samples	Variables	Var./Samp.	Classes
Wine (UCI)	177	13	0.17	3
Ionosphere (UCI)	351	35	0.01	2
Breast Cancer (UCI)	569	32	0.056	2
Lymphoma	96	4026	41.9	2
Colon Cancer	62	2000	32	2

All the datasets have been standardised subtracting the median and dividing by the inter-quantile range.

For high dimensional problems such as gene expression datasets, dimension reduction helps to improve significantly the clustering results [9]. Therefore, for the algorithms based on a single dissimilarity we have considered different number of genes 280, 146, 101, 56 and 34 obtained by feature selection [11]. Genes have been ranked according to the method proposed by [4]. Then, we have chosen the subset that gives rise to the smallest error. Considering a larger number of genes or even the whole set of genes does not help to improve the clustering performance. Regarding the algorithm proposed to integrate several dissimilarities, we have considered all the dissimilarities obtained for the whole set of dimensions.

The similarity/dissimilarity constraints are obtained as in [17]. $\mathcal{S}$ is generated by picking a random subset of all pairs of points sharing the same class label. The size is chosen such that the number of connected components is roughly 20% of the size of the original dataset. $\mathcal{D}$ is chosen in a similar way although the size in this case is less relevant.

Regarding the value of the parameters, the number of clusters is set up to the number of classes, C_S and C_D are regularization parameters and the optimal value is determined by cross-validation over the subset of labeled patterns. Finally, kernel k -means is restarted randomly 20 times and the errors reported are averages over 20 independent trials.

Clustering results have been evaluated considering two objective measures. The first one is the accuracy. It determines the probability that the clustering

Table 2. Accuracy for k -means clustering considering different dissimilarities. The results are averaged over twenty independent random subsets $\mathcal{S}$ and $\mathcal{D}$.

Technique	Kernel	Wine	Ionosphere	Breast	Colon	Lymphoma
k -means (Euclidean)	linear	0.92	0.72	0.88	0.87	0.90
	pol. 3	0.87	0.73	0.88	0.88	0.90
k -means (Best diss.)	linear	0.94	0.88	0.90	0.88	0.94
	pol. 3	0.94	0.88	0.90	0.88	0.93
		χ^2	Maha.	Manha.	Corr./euclid.	χ^2
Comb. dissimilarities	linear	0.94	0.90	0.92	0.89	0.95
	pol. 3	0.96	0.89	0.92	0.90	0.92
Metric learning (Xing)	linear	0.87	0.74	0.85	0.87	0.90
	pol. 3	0.51	0.74	0.86	0.88	0.90

Table 3. Adjusted RandIndex for k -means clustering considering different dissimilarities. The results are averaged over twenty independent random subsets $\mathcal{S}$ and $\mathcal{D}$.

Technique	Kernel	Wine	Ionosphere	Breast	Colon	Lymphoma
k -means (Euclidean)	linear	0.79	0.20	0.59	0.59	0.65
	pol. 3	0.67	0.21	0.60	0.59	0.65
k -means (Best diss.)	linear	0.82	0.58	0.66	0.59	0.77
	pol. 3	0.81	0.58	0.66	0.59	0.76
		χ^2	Maha.	Manha.	Corr./euclid.	χ^2
Comb. dissimilarities	linear	0.82	0.63	0.69	0.60	0.79
	pol. 3	0.85	0.60	0.69	0.63	0.73
Metric learning (Xing)	linear	0.68	0.23	0.50	0.54	0.66
	pol. 3	0.50	0.23	0.52	0.58	0.65

agrees with the “true” clustering in the sense that the pair of patterns belong to the same or different clusters. It has been defined as in [17]:

$$\text{accuracy} = \sum_{i>j} \frac{1\{1\{c_i = c_j\} = 1\{\hat{c}_i = \hat{c}_j\}\}}{0.5m(m-1)}, \quad (21)$$

where c_i is the true cluster label for pattern x_i , $\hat{c}_i$ is the corresponding label returned by the clustering algorithm and m is the number of patterns. One problem of the accuracy is that the expected value for two random partitions is not zero. Therefore, we have computed also the adjusted randindex defined in [6] that avoids this problem. This index is also normalized between zero and one and larger values suggest better clustering.

Tables 2 and 3 show the accuracy and the adjusted randindex for the clustering algorithms evaluated. We have compared with a standard metric learning strategy proposed by [17], k -means clustering based on the Euclidean distance and k -means considering the best dissimilarity out of ten different measures. Both tables indicates which is the best distance for each case.

From the analysis of tables 2 and 3, the following conclusions can be drawn:

- The combination of dissimilarities improves significantly a standard metric learning algorithm for all the datasets considered. Our method is robust to overfitting and outperforms the algorithm proposed by Xing [17] in high dimensional datasets such as Colon cancer and Lymphoma. These datasets exhibit a high level of noise. We can explain this because the algorithm based on a combination of dissimilarities allows to integrate distances computed for several dimensions discarding the noise and reducing the errors.
- The combination of dissimilarities improves usually kernel k -means based solely on the best dissimilarity. This suggests that the integration of several dissimilarities allows to extract complementary information that may help to improve the performance. Besides, the algorithm proposed always achieves at least the same performance that k -means based on the best dissimilarity. Only for lymphoma and polynomial kernel we get worst results, probably because the value assigned to the regularization parameters overfit the data. We remark that the algorithm proposed, helps to overcome the problem of choosing the best dissimilarity, the kernel and the optimal dimension. This is a quite complex and time consuming task for certain applications such as Bioinformatics.

Finally, the combination of dissimilarities improves always the standard k -means clustering based on the Euclidean measure.

- Tables 2 and 3 show that the best distance depends on the dataset considered. Moreover, we report that the performance of k -means depends strongly on the particular measure employed to evaluate the sample proximities.

Figure 1 shows a boxplot diagram for the accuracy and adjusted randindex coefficients. Odds numbers correspond to the combination of dissimilarities and the even ones to the metric learning algorithm proposed by Xing. We can see that the differences between the method proposed here and the one proposed

Fig. 1. Boxplots that compare the combination of dissimilarities with the metric learning algorithm proposed by Xing according to (a) accuracy and (b) Adjusted RandIndex. All the boxplots consider linear kernels.

by Xing are statistically significant at 95% confidence level for all the datasets considered.

5 Conclusions

In this paper, we propose a semi-supervised algorithm to learn a combination of dissimilarities from equivalence constraints. The error function includes a penalty term that controls the complexity of the family of distances considered and the optimization is based on a robust quadratic programming approach that does not suffer from the problem of local minima.

The experimental results suggest that the combination of dissimilarities improves almost always the performance of clustering algorithms based solely on a single dissimilarity. Besides, the algorithm proposed improves significantly a standard metric learning algorithm for all the datasets considered in this paper and is robust to overfitting.

Future research trends will focus on the application of this formalism to the integration of heterogeneous data sources.

Appendix A

This appendix introduces shortly the Empirical Kernel Map that allow us to work with non-Euclidean dissimilarities considering kernel methods [12].

Let $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a dissimilarity and $R = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ a subset of representatives drawn from the training set. Define the mapping $\phi: \mathcal{F} \rightarrow \mathbb{R}^n$ as:

$$\phi(z) = D(z, R) = [d(z, \mathbf{x}_1), d(z, \mathbf{x}_2), \dots, d(z, \mathbf{x}_n)] \quad (22)$$

This mapping defines a dissimilarity space where feature i is given by $d(\cdot, \mathbf{x}_i)$. The kernel of dissimilarities can be defined as the dot product of two dissimilarity vectors in feature space.

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \sum_{i=1}^n d(\mathbf{x}, p_i) d(\mathbf{x}', p_i) \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \quad (23)$$

References

1. Cox, T.F., Cox, M.A.A.: Multidimensional scaling, 2nd edn. Chapman & Hall/CRC, USA (2001)
2. Cristianini, N., Kandola, J., Elisseeff, J., Shawe-Taylor, A.: On the kernel target alignment. *Journal of Machine Learning Research* 1, 1–31 (2002)
3. Shawe-Taylor, J., Cristianini, N.: *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge (2004)
4. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association* 97(457), 77–87 (2002)

5. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a Mahalanobis Metric from Equivalence Constraints. *Journal of Machine Learning Research* 6, 937–965 (2005)
6. Hubert, L., Arabie, P.: Comparing Partitions. *Journal of Classification*, 193–218 (1985)
7. Huang, D., Pan, W.: Incorporating Biological Knowledge into Distance-Based Clustering Analysis of Microarray Gene Expression Data. *Bioinformatics* 22(10), 1259–1268 (2006)
8. Kwok, J.T., Tsang, I.W.: Learning with Idealized Kernels. In: *Proceedings of the Twentieth International Conference on Machine Learning*, Washington DC, pp. 400–407 (2003)
9. Jeffery, I.B., Higgins, D.G., Culhane, A.C.: Comparison and Evaluation Methods for Generating Differentially Expressed Gene List from Microarray Data. *BMC Bioinformatics* 7(359), 1–16 (2006)
10. Lanckriet, G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.: Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* 3, 27–72 (2004)
11. Martín-Merino, M., Blanco, A., De Las Rivas, J.: Combining Dissimilarities in a Hyper Reproducing Kernel Hilbert Space for Complex Human Cancer Prediction. *Journal of Biomedicine and Biotechnology*, 1–9 (2009)
12. Pekalska, E., Paclick, P., Duin, R.: A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research* 2, 175–211 (2001)
13. Soon Ong, C., Smola, A., Williamson, R.: Learning the kernel with hyperkernels. *Journal of Machine Learning Research* 6, 1043–1071 (2005)
14. Vapnik, V.: *Statistical Learning Theory*. John Wiley & Sons, New York (1998)
15. Wu, G., Chang, E.Y., Panda, N.: Formulating distance functions via the kernel trick. In: *ACM SIGKDD*, Chicago, pp. 703–709 (2005)
16. Zhao, B., Kwok, J.T., Zhang, C.: Multiple Kernel Clustering. In: *Proceedings of the Ninth SIAM International Conference on Data Mining*, Nevada, pp. 638–649 (2009)
17. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance Metric Learning, with Application to Clustering with Side-Information. In: *Advances in Neural Information Processing Systems*, vol. 15, pp. 505–512. MIT Press, Cambridge (2003)
18. Xiong, H., Chen, X.-W.: Kernel-Based Distance Metric Learning for Microarray Data Classification. *BMC Bioinformatics* 7(299), 1–11 (2006)