

Microarray Design using the Hilbert–Schmidt Independence Criterion

Justin Bedo

The Australian National University,
NICTA, and the University of Melbourne

Abstract. This paper explores the design problem of selecting a small subset of clones from a large pool for creation of a microarray plate. A new kernel based unsupervised feature selection method using the Hilbert–Schmidt independence criterion (HSIC) is presented and evaluated on three microarray datasets: the Alon colon cancer dataset, the van 't Veer breast cancer dataset, and a multiclass cancer of unknown primary dataset. The experiments show that subsets selected by the HSIC resulted in *equivalent or better* performance than supervised feature selection, with the added benefit that the subsets are not target specific.

1 Introduction

Feature selection is an important procedure in data mining. The elimination of features leads to smaller and more interpretable models and can improve generalisation performance. Supervised methods produce feature subsets tailored towards the prediction target and are applicable when labels are available. In contrast, unsupervised methods select features that capture some of the information contained within the whole dataset without requiring labels; as no labels are used the feature selections are not target specific.

The problem of designing a sugarcane microarray plate by choosing a subset of approximately 7000 clones from an initial pool of 50,000 clones is studied herein. As the initial pool of clones contains some highly correlated pairs, there is a preference towards *decorrelation*. Furthermore, the array must remain as general as possible and not be tailored towards any specific phenotypes. As such, this is an *unsupervised* selection problem.

The *Hilbert–Schmidt independence criterion* [1] (HSIC) HSIC is a dependence measure between two random variables which is closely related to *kernel target alignment* [2] and *maximum mean discrepancy* [3] (MMD). Previous papers [4,5] used the HSIC for supervised feature selection and demonstrated that the method had good performance and flexibility on several genomics datasets. This paper presents an unsupervised variant named *unsupervised feature selection by the HSIC* (UBHSIC, pronounced [ˈu.bə-sik]).

UBHSIC was evaluated in several experiments comparing the selection using various kernels to supervised feature selection. As labels are not available on the sugarcane dataset, UBHSIC was evaluated on three cancer genomics datasets: the

Alon colon cancer dataset [6,7,8,9], the van 't Veer breast cancer dataset [10], and a multiclass cancer of unknown primary (CUP) dataset [11]. The CUP dataset closely resembles the sugarcane problem as it was intended for the development of a clinical test on a lower resolution platform.

2 HSIC and UBHSIC

The HSIC is a quantity that measures the mutual dependence between two variables. For the task of unsupervised feature selection, the dependence between subsets of features and the full set of features is measured by the HSIC; a subset with *maximum dependence* on the full dataset is desired. This section gives an overview of the HSIC and specifies the unsupervised feature selection problem as a constrained optimisation problem.

Let

$$X := \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

be a finite dataset in matrix form with $x_{ij} \in \mathbb{R}$, where n is the number of samples and m is the number of features. Each row \mathbf{x}_i corresponds to a sample, and each column \mathbf{x}_j corresponds to a feature.

Let $\theta \in 2^m$ be a subset of features, where 2^m denotes the power set of $\{1, \dots, m\}$, and define X_θ as the dataset *restricted to only the features* in θ , i.e., the features with indices not in θ are discarded. By this definition, X_θ is a matrix with dimension $n \times |\theta|$, where $|\cdot|$ denotes set cardinality. The dependence between the reduced dataset X_θ and the full dataset X is the quantity we wish to maximise. The HSIC measures this dependence through *kernel functions* [12,13].

A kernel function defines the inner product between two points of a Hilbert space, and can be considered intuitively as a measure of similarity. Indeed, the correlation function $\text{cor}(\mathbf{x}, \mathbf{x}') := \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|\mathbf{x}\| \|\mathbf{x}'\|}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, is a kernel function used in the experiments section. Given a kernel function k , the *kernel matrix* is defined $[K_{ij}]_{1 \leq i, j \leq n} := k(\mathbf{x}_i, \mathbf{x}_j)$. The kernel matrix of the full dataset X is referred to as K , and the kernel matrix of the reduced dataset X_θ as K_θ .

An estimator for the HSIC using these two kernel matrices [1] is

$$\text{tr}(K_\theta H K H), \tag{1}$$

where tr is the *matrix trace* (the sum over the elements of the main diagonal), $H := Id - \frac{1}{n} Id$, Id is the identity matrix, and the subtraction is *element wise*. Using this dependence measure, the unsupervised selection task can simply be stated as

$$\begin{aligned} \max_{\theta} \text{tr}(K_\theta H K H) & \tag{2} \\ \text{such that} & \\ |\theta| = m' & \end{aligned}$$

for some $1 \leq m' < m$. Solving this optimisation equation for a set θ gives the UBHSIC solution.

The solution to the optimisation equation is explicit in the linear kernel case where $K := XX^T$. Let $M := HKH$. The HSIC estimator is then

$$\begin{aligned} \text{tr}(KM) &= \text{tr}(XX^T M) \\ &= \text{tr}(X^T M X) \\ &= \sum_j \mathbf{x}_j^T M \mathbf{x}_j. \end{aligned}$$

Thus, in the case of a linear kernel the features are *independent* and can be ranked by $\mathbf{x}_j^T M \mathbf{x}_j$ and greedily selected.

For other kernels, an analytical solution does not exist and a good subset must be found through searching. The forward selection and recursive elimination greedy nested subset strategies [14] can be used to find a good solution if the number of features is not large. This approach was used for the supervised variant presented by Song *et al.* [4,5]. Alternatively, a good solution can be found using combinatorial optimisation algorithms such as simulated annealing. For the sugarcane dataset, a good solution to (2) is found using an annealing algorithm as nested subset selection is unattractive due to the large pool of initial features. Efficient solving the optimisation problem is an open problem.

3 Results and Discussion

The proposed method was analysed on several cancer genomics datasets using different kernels. These kernels are defined as follows:

Radial basis function (RBF): $k(\mathbf{x}, \mathbf{x}') := \exp(-\sigma \|\mathbf{x} - \mathbf{x}'\|_2^2)$ with σ set as the inverse median of the squared distances $\|\mathbf{x} - \mathbf{x}'\|_2^2$ between points in the dataset

Linear: $k(\mathbf{x}, \mathbf{x}') := \langle \mathbf{x}, \mathbf{x}' \rangle$

Polynomial: $k(\mathbf{x}, \mathbf{x}') := (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^d$ for $d \in \{2, 3\}$

Variance: $k(\mathbf{x}, \mathbf{x}') := \frac{\langle \mathbf{x}, \mathbf{x}' \rangle^2}{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{x}', \mathbf{x}' \rangle}$

The variance kernel was chosen to produce highly decorrelated selections. The preference towards decorrelation is indirectly encoded as $\langle \mathbf{x}, \mathbf{x}' \rangle / \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{x}', \mathbf{x}' \rangle}$ is the cosine of the angle between the two vectors \mathbf{x} and \mathbf{x}' . Thus, as adding a feature highly correlated with another already selected feature will not affect the angle between the vectors as much as a feature orthogonal to all selected features, one may postulate that the kernel used with UBHSIC will produce highly decorrelated selections.

Three cancer genomics datasets were analysed, the van 't Veer breast cancer dataset [10] and a colon cancer dataset [6,7,8,9]. The van 't Veer dataset consists of 98 samples, 46 with a distant metastasis and 52 with no metastasis. Each sample has 5952 dimensions. The colon cancer dataset has 62 samples, 22 normal

and 40 cancerous, and 2000 dimensions per sample. Both datasets are 2-class classification problems.

The final cancer genomics dataset is a cancer of unknown primary (CUP) dataset [11]. This is a multiclass classification dataset where the aim is to develop a predictor for the site of origin of a tumour from a microarray of a sample. The dataset consists of 14 classes, 220 samples, and 9630 features. Not each class is represented equally, with the smallest class containing only 3 samples and the largest containing 34.

To gauge the utility of feature subsets selected by UBHSIC for prediction, the reduced datasets were evaluated using supervised classification and generalisation estimation. The performance achievable from the reduced datasets were also compared to a fully supervised selection approach.

The classification and supervised feature selection algorithm used was a centroid based classifier and supervised feature selector [15]. This method was chosen as it is simple, fast, and has performed well on these particular datasets [15]. For the multiclass CUP dataset, a one-vs-all architecture [16] was used in conjunction with the centroid classifier to produce a multiclass classifier. For generalisation performance estimation, the ϵ -0 bootstrap estimator [17] was used with 200 repetitions. The *area under the ROC curve* (AROC) [18] was used as a performance metric for the two-class datasets. A multiclass extension to the AROC was used [19] for the CUP dataset.

Each dataset was analysed by applying UBHSIC with the various kernels to reduce the full dataset. The centroid classifier and supervised feature selector was then applied to the UBHSIC reduced datasets to evaluate the performance. The same centroid classifier and supervised feature selector was applied to the full dataset to obtain the performance achievable using supervised selection only without any UBHSIC pre-filtering.

Figure 1 shows the results of pre-filtering using UBHSIC down to 50 (Subfigure 1a) and 500 features (Subfigure 1b) on the van 't Veer dataset. With the reduction to 500 features, the linear, RBF and variance kernels do very well; they achieve a level of performance equivalent to the full dataset at higher numbers of features and exceed the performance at lower numbers of features. The two polynomial kernels initially perform poorly, but after mild supervised feature selection the performance equals that of the other kernels and the full dataset. Under aggressive reduction down to 50 features, somewhat surprising results are obtained; the maximum performance achieved was *substantially better* than the full dataset using a polynomial kernel of degree 2 despite the operating with only 32 features. Furthermore, the variance kernel achieves very high performance at the eight features operating point. Both are significantly fewer than the original 70 genes proposed for classification by the original paper [10].

Performing the same experiments on the colon cancer dataset yielded the results in Figure 2. Again, strong performance when using the variance and RBF kernels is observable in Subfigure 2b; RBF produced very good results after further supervised filtering down to a few features (4) while the variance kernel produced very similar results to the full dataset. The linear and polynomial

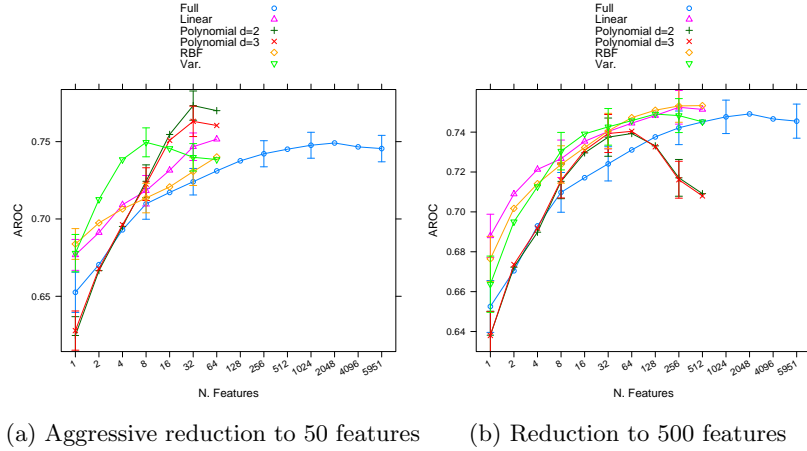


Fig. 1: van 't Veer dataset with centroid classifier and feature selector. Results are using the ϵ -0 bootstrap with 200 repetitions. Error bars show 95% confidence interval. Subfigure (a) shows the performance of the dataset reduced to 50 features using the UBHSIC procedure and various kernels. Each plot corresponds to a different kernel, with the purple plot corresponding to the CFS-centroid method on the entire dataset (i.e., without prefiltering using UBHSIC). The 5 plots where prefiltering using UBHSIC was used do not extend above 50 features, and further supervised filtering using the CFS was applied to determine the maximum performance achievable from the reduced datasets. Subfigure (b) is similar to subfigure a, except with less aggressive UBHSIC reduction (reduced to 500 features instead of 50).

kernels do not perform well on this dataset; this is supported by the results shown in Subfigure 2a where the linear and polynomial kernels again perform poorly, but the RBF and variance kernels perform well.

Finally, the results of applying the unsupervised feature selection to the CUP dataset is shown in Figure 3. As this dataset is a larger dataset (220 samples) than both the colon and van 't Veer datasets, a less aggressive filtering was applied. Subfigure 3b shows the performance curves obtained after filtering to 500 features. At 500 features, the variance kernel produces a subset with equivalent performance to the full dataset. At the aggressive reduction to 100 features, the performance does not suffer greatly for the variance kernel. The other kernels do not perform well on this dataset.

Furthermore, the 500 feature subset selected by the variance kernel outperformed the full dataset at low numbers of features. The performance achieved below 32 features is greater than the performance at the same operating point obtained with the full dataset. Given this performance, a satisfactory operating

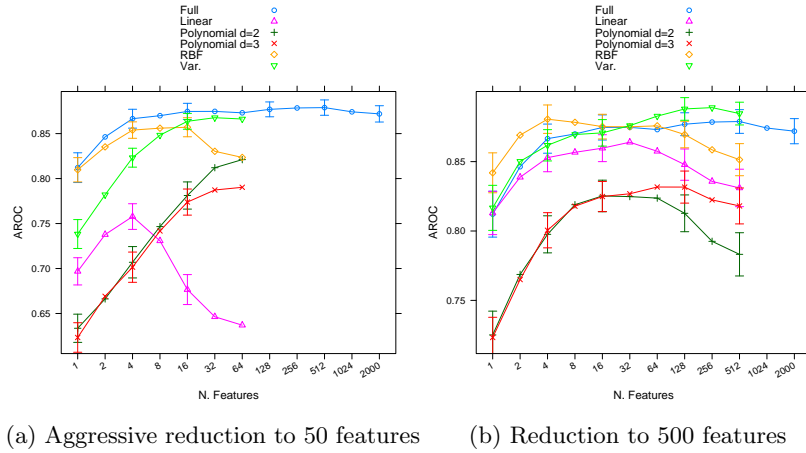


Fig. 2: Colon cancer dataset with centroid classifier and feature selector. $\epsilon=0$ bootstrap with 200 repetitions. Error bars show 95% confidence interval. The experiment is identical to Figure 1, except with a different dataset.

point at 16 features or even 8 features per class may be chosen, resulting in a very sparse predictor.

In summary, these results show that unsupervised pre-filtering does not degrade the classification performance and can actually improve the performance at few features. The RBF and variance kernels perform very well across both two-class datasets, with the other kernels not performing as consistently. On the multiclass dataset, the variance kernel is the only kernel that performed well. The aggressive feature reduction down to 50 features for the two-class datasets and 100 features for the CUP dataset showed surprisingly good performance, suggesting that the full datasets contains significant redundancy and can be highly compressed without significant loss of performance.

3.1 van 't Veer in detail

To gain a better understanding of the relation between features selected by UBHSIC, the feature subsets obtained on the van 't Veer data were visualised. Subfigure 4a shows the full unfiltered dataset projected down onto the first two principal components with each sample represented by a number. It is clear from the projection that sample 10 is an outlier, sitting far away from the other samples. Excluding this sample and reprojecting the data obtains the embedding shown in Subfigure 4b. Here one can observe that the samples roughly form two groups separated mostly by the first principal component (x -axis).

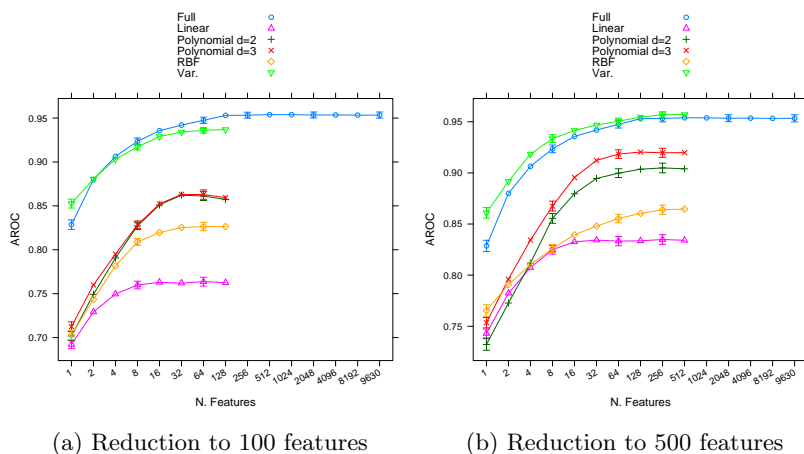


Fig. 3: CUP cancer dataset with centroid classifier and feature selector. $\epsilon=0$ bootstrap with 200 repetitions. Error bars show 95% confidence interval. Number of features shown is per class not overall. Experiment details are as in Figure 1.

Subfigure 5a displays a biplot [20] of the dataset filtered down to 100 features using the linear kernel and UBHSIC. In the figure, samples are shown as black points and features as red vectors. If two feature vectors have a small angle then they are highly correlated. From the figure the two-group structure observable on the original projection (Subfigure 4b) is maintained. Furthermore, the selected features are strongly positioned along the first principal component. This is not unexpected as a linear kernel is expected to favour the first principal component, and as features are selected independently it is also expected to select highly correlated feature sets. Indeed, a selection of 100 features most correlated with the first principal component yields a subset of features with 77 features in common with the subset selected by the linear kernel and UBHSIC.

The biplot produced using the RBF kernel (Figure 6) resembles the linear kernel results in that the two-group structure is preserved with many features selected along the first principal component. However, in comparison the features are more spread out in two fan-like structures, each spanning one of the groups well, whereas the “fans” formed by the linear kernel are not as spread out and well aligned with the groups. The RBF kernel is selecting sets with high cross-correlation; this is evident from the number of feature vectors with small interior angles.

Running the same analysis using the polynomial filter of degree 2 yields the results shown in Figure 6. Interestingly, the selected feature subset appears to have generated an outlier that is clearly visible in Subfigure 6a; removing this outlier produces a vastly different projection as shown in Subfigure 6b. In this figure the feature vectors can be observed to have a “radial” pattern,

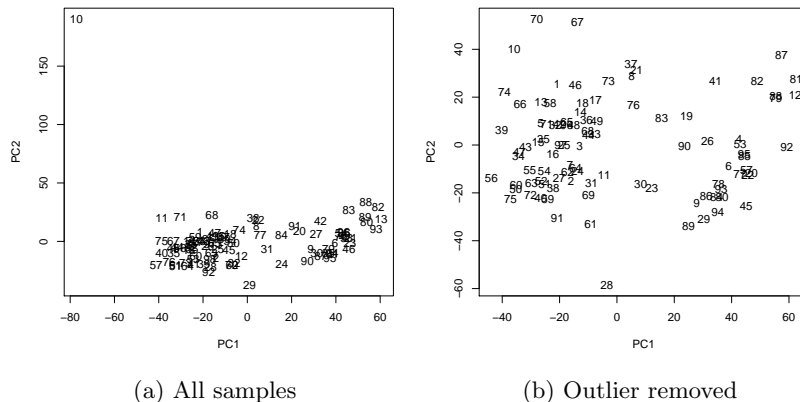


Fig. 4: Biplot of samples and features projected onto first two principal components using the full van ’t Veer dataset. The x -axis is the first principal component, and the y -axis is the second. The sample marked as 10 in subfigure a is clearly an outlier; removing the outlier and reprojecting the samples produces the embedding shown in subfigure b.

indicating the selected features do not have high cross-correlation. Similar results are obtained with the polynomial kernel of degree 3 (not shown). The indication here is that polynomial kernels tend to favour feature subsets with lower cross-correlation than the RBF and linear kernels.

Finally, the variance kernel is shown in Figure 7. Unlike the polynomial kernel, the variance kernel did not produce any new outliers and resulted in a more “radial” pattern than the polynomial filter. This indicates that the selected features are highly decorrelated as postulated previously.

These results indicate the linear and RBF kernels produce subsets with high cross-correlations; the linear kernel is especially highly cross-correlated and aligned with the first principal component while the RBF kernel spans the samples well and is less cross-correlated. The polynomial kernel and variance kernels result in much more decorrelated results, with the variance kernel producing highly decorrelated selections. Given the classification performance observed on the van ’t Veer datasets, the RBF and variance kernels are both good choices and can be selected depending if one wishes to obtain whitened data or not.

4 Conclusions

A method for unsupervised feature selection, UBHSIC, was presented and evaluated on several bioinformatics datasets. The results were very promising: on the

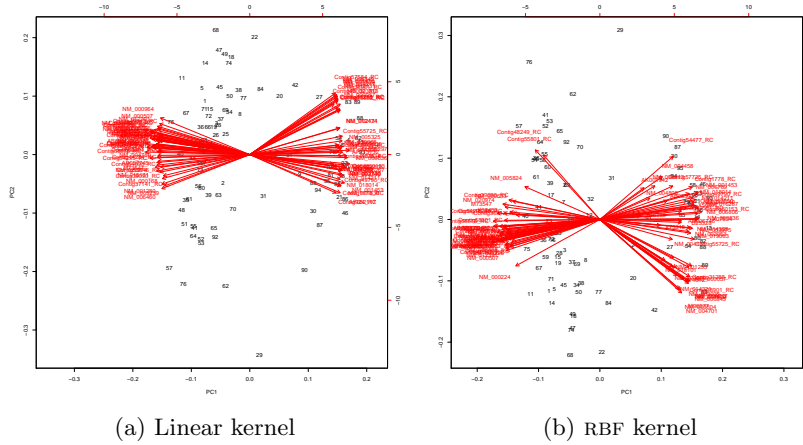


Fig. 5: Biplot after filtering the van 't Veer dataset down to 100 features using the linear and RBF kernels. Both kernels produce selections polarised along the first principal component, though the RBF kernel selections span the samples better than the linear kernel selections.

cancer genomics datasets the classification performance after pre-filtering using UBHSIC was equivalent or better than the performance obtained using the full dataset. The RBF and variance kernels show good performance on all two-class datasets, and the variance kernel showed good performance on the multiclass dataset. Furthermore, the variance kernel producing highly decorrelated selections as postulated.

The high level of classification performance observed after filtering strongly suggests shifting to a lower resolution platform by selecting a subset of clones using the presented method is a viable option. In particular, UBHSIC may be a reasonable solution to the inspiring sugarcane microarray plate design problem. Furthermore, the feature subsets obtained using UBHSIC procedure are not tailored for a specific target and thus may be used to predict many different phenotypes, though further supervised feature selection may be needed to reach the maximum performance.

Acknowledgment

I acknowledge the permission of NICTA to publish this paper. NICTA is funded by the Australian Governments Department of Communications, Information Technology and the Arts and the Australian Council through Backing Australias Ability and the ICT Centre of Excellence programs. I thank Adam Kowalczyk, Geoff Macintyre, and Alex Smola for helpful discussions and help in preparing this manuscript.

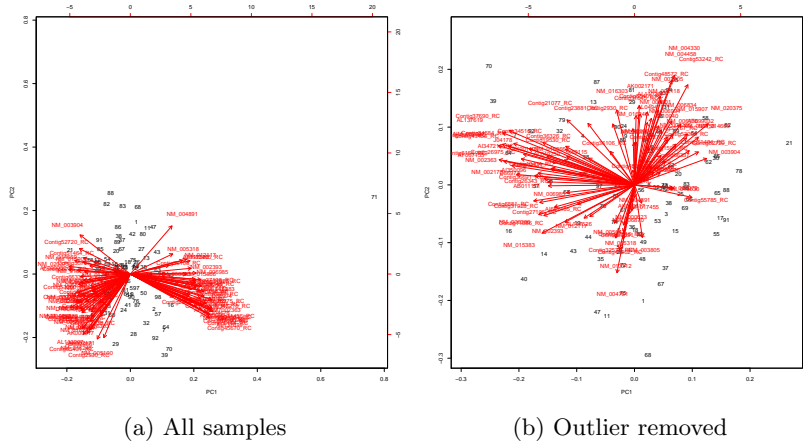


Fig. 6: Biplot after filtering the van 't Veer dataset down to 100 features using a polynomial kernel of degree 2. Unlike the linear and RBF kernels, the pattern is more radial, suggesting the selection has less coregulation. With this selection, an outlier is apparent in subfigure (a). Subfigure (b) shows the biplot with the outlier removed.

References

1. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. *Algorithmic Learning Theory: 16th International Conference* (Jan 2005)
2. Cristianini, N., Shawe-Taylor, J.: On kernel-target alignment. *Neural Information Processing Systems* **14** (Jan 2002)
3. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems* (Jan 2007)
4. Song, L., Bedo, J., Borgwardt, K., Gretton, A., Smola, A.: Gene selection via the bahsic family of algorithms. *Bioinformatics* (Jan 2007)
5. Song, L., Smola, A., Gretton, A., Borgwardt, K., Bedo, J.: Supervised feature selection via dependence estimation. *Proceedings of the 24th international conference on Machine Learning* (Jan 2007)
6. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci US A* (Jan 1999)
7. Ambrose, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* **99**(10) (May 2002) 6562–6
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. *Machine Learning* **46** (Nov 2002) 389–422

- bioinformatics. Proceedings of the Australian Joint Conference on Artificial Intelligence (2006)
16. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. *The Journal of Machine Learning Research* (Jan 2004)
 17. Efron, B.: How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* (Jan 1986)
 18. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143**(1) (Apr 1982) 29–36
 19. Hand, D., Till, R.: A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning* (Jan 2001)
 20. Gabriel, K.: The biplot graphic display of matrices with application to principal component analysis. *Biometrika* (Jan 1971)