An Empirical Comparison of Feature Reduction Methods in the Context of Microarray Data Classification

Hans A. Kestler[1,2] and Christoph Müssel[2]

[1] Department of Neural Information Processing, University of Ulm, 89069 Ulm, Germany
hans.kestler@uni-ulm.de
[2] Department of Internal Medicine I, University Hospital Ulm, Robert-Koch-Str. 8, 89081 Ulm, Germany
christoph.muessel@uni-ulm.de

Abstract. The differentiation between cancerous and benign processes in the body often poses a difficult diagnostic problem in the clinical setting while being of major importance for the treatment of patients. Measuring the expression of a large number of genes with DNA microarrays may serve this purpose. While the expression level of several thousands of genes can be measured in a single experiment, only a few dozens of experiments are normally carried out, leading to data sets of very high dimensionality and low cardinality. In this situation, feature reduction techniques capable of reducing the dimensionality of data are essential for building predictive tools based on classification.

Methods and Data: We compare the popular feature selection and classification method PAM (Tibshirani et al.) to several other methods. Feature reduction and feature ranking methods, such as Random Projection, Random Feature Selection, Area under the ROC curve and PCA are applied. We employ these together with the classification component of PAM, Linear Discriminant Analysis (LDA), a Nearest Prototype (NP) classifier and linear support vector machines (SVMs). We apply these methods to three publicly available linearly separable gene expression data sets of varying cardinality and dimensionality.

Results and Conclusions: In our experiments with the gene expression data we could not discover a clearly superior algorithm, instead most surprisingly we found that feature reduction using random projections or selections performed often equally well.

1 Background

The differentiation between cancerous and benign (non-cancerous) processes in the body often poses a difficult diagnostic problem in the clinical setting while being of major importance for the treatment of patients. Since cancer development is thought to be caused by the accumulation of complex genetic alterations in the affected cells and tissues, the differentiation of cancerous vs. non-cancerous

F. Schwenker and S. Marinai (Eds.): ANNPR 2006, LNAI 4087, pp. 260–273, 2006.

clinical samples is an important application –together with feature reduction methods– for the interpretation of DNA array data. Gene expression data have two dimensions, genes on one side, and measurements on the other side. While the expression level of several thousands of genes can be measured in a single experiment, only a few dozens of experiments are normally carried out, leading to data sets of very high dimensionality and low cardinality. In this situation, feature reduction techniques capable of reducing the dimensionality of data are essential for building predictive tools based on classification. It is thought that only a small fraction of the features are needed for classification, while most of the features are not only irrelevant, but may even disturb the classification. This poses the problem of reducing the feature set. The feature reduction methods can be divided into two classes: Transformation methods, which project the original feature space into a lower-dimensional space, and feature selection methods, which choose a subset of the original features. The former have the advantage of not throwing away any features completely, whereas the latter provide results that can be interpreted more easily. One popular feature selection method is the use of shrunken centroids proposed by Tibshirani et al. [1, 2], which is also known as Prediction Analysis for Microarrays (PAM).

We compare the popular feature selection and classification method PAM [1,2] to several other methods. Feature reduction and feature ranking procedures, such as Random Projection, Random Feature Selection, Area under the ROC curve and PCA are applied [3, 4, 5]. We employ these together with the classification component of PAM, Linear Discriminant Analysis (LDA), a Nearest Prototype (NP) classifier and linear support vector machines (SVMs) [4, 6].

2 Methods

This section consists of a brief description of the used feature reduction and classification methods (for details see the cited references and any standard text such as Duda&Hart or Webb [4, 5]), and the employed gene expression data sets and testing procedures.

2.1 Feature Reduction Methods

Prediction Analysis for Microarrays (PAM) was described by Tibshirani et al. in [1, 2]. The PAM algorithm performs both feature reduction and classification. PAM chooses class representatives (prototypes, centroids) for every feature and moves (shrinks) them towards the overall centroid (not taking into account any class information) of that particular feature using a fixed threshold value. Whenever a class centroid has zero distance to the feature centroid, it does not play a role in the classification any more and can be discarded. If all class centroids in a feature have been discarded, the feature itself is removed. PAM uses exactly one representative vector per class. The components of the representatives are the centroids of the class samples in each feature.

With samples $j = 1, \ldots, n$, classes $1, \ldots, K$ and $i = 1, \ldots, p$ features/genes, the initial ith component of the centroid of class k is $\overline{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k}$, where

C_k are the indices of the samples in class k. The ith component of the overall centroid of feature i is $\overline{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$.

The standardized distance between the class centroid and the overall centroid in feature i is

$$d_{ik} = \frac{\overline{x}_{ik} - \overline{x}_i}{m_k \cdot (s_i + s_0)}$$

where

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ij} - \overline{x}_{ik})^2,$$

and $s_0 = \text{median}\{s_i\}$.

In [1], m_k was defined as $m_k = \sqrt{1/n_k + 1/n}$. However, the PAM implementation provided in the *pamr* package [7] defines it as $m_k = \sqrt{1/n_k - 1/n}$. The second definition was used in the experiments.

The shrunken centroid is

$$\overline{x}'_{ik} = \overline{x} + m_k(s_i + s_0)d'_{ik}$$

where

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$$

($t_+ = t$ if $t > 0$ and zero otherwise)

This means that all centroids whose distances to the overall centroid of the feature are less than Δ will be the same as the overall centroid and can therefore be eliminated. Tibshirani et al. [1, 2] propose using a sequence of numbers as threshold values Δ and finding the best one by a 10-fold cross-validation on the training data. We used this unmodified version of PAM in the nested cross-validation tests, but employed a different method for the cross-validation runs with fixed numbers of features:

In fact, choosing arbitrary sequential threshold values Δ is an inaccurate method, as all possible thresholds can be calculated. The choice of the threshold values should have an impact on the number of remaining centroids, and consequently the set of reasonable thresholds consists of the d_{ik}s. Instead of the sequential threshold procedure, we used those pre-calculated threshold values. In addition, it is necessary for our experiments to fix the number of remaining features in order to be comparable to other feature reduction methods. Therefore, we did not choose the best threshold using a cross-validation, but utilized the threshold that leaves exactly p' features:

1. Calculate $\max_k |d_{ik}|$ for each feature i
2. Sort these thresholds in descending order
3. Pick the $(p'+1)$th threshold

The PAM classification rule is described in Sect. 2.2.

Random Feature Selection (RF) simply chooses a specified number of features at random. Let I be the set of feature indices $\{1, \ldots, p\}$. Then $I' \subset I$ is a random sample of p' indices drawn from I uniformly and without replacement,

where p' is the desired number of reduced features. If we let D be the $p \times s$ matrix containing the data (with p being the original number of dimensions), the feature reduction is performed by picking all rows of D with indices $i \in I'$. Random Feature Selection is very simple, but the usefulness of the selected features is not evaluated and features are randomly discarded that may contain important information. It serves as a baseline for comparing the algorithms.

A Random Projection (RP) is a transformation with a matrix R whose entries r are chosen from a distribution that is symmetric about the origin. The idea of Random Projections is based on a lemma due to Johnson and Lindenstrauss [8]. The lemma basically states that any set of n points in $\mathbf{R}^d$ can be projected into $\mathbf{R}^k, k \geq O(\epsilon^{-2} \log n)$ so that all distances are preserved up to a factor of $1 \pm \epsilon$. Vempala [3] further describes the distributions that can be used for generating the matrix entries. For the experiment, p' vectors R_i of length p were chosen from the standard normal distribution $N(0, 1)$, where p is the original number of features and p' is the desired number of reduced features. If we let R be the $p \times p'$ matrix whose columns are the vectors $R_1 \ldots R_{p'}$ and D the $p \times s$ matrix containing the data (with s being the number of samples), the projection is $D' = R^T \cdot D$.

Principal Component Analysis (PCA) is a linear transformation that aligns the first axis (the principal component) of the coordinate system along the greatest variance. Formally, given a $p \times s$ data matrix D (where p is the number of features and s is the number of samples), we use PCA for dimensionality reduction to p' features, i.e. the eigenvectors of the covariance matrix are sorted in decreasing order and we use the first p' eigenvectors.

The Area Under the ROC Curve can be a measure of classification quality of a feature in the two class scenario. A ROC curve has the 1-specificity on the horizontal axis and the sensitivity on the vertical axis. It visualizes the possibilities of separating the classes and allows to adjust the misclassification rates for both classes separately. The area under the ROC curve (AUC) is a measurement of discrimination. The closer it approaches 1, the better is the feature suitable for classification. Feature reduction is done by using only first p' features after sorting them in decreasing order according to the AUC.

2.2 Classifiers

The PAM Classifier described in [1] has the following discriminant function:

$$\delta_k^{\mathrm{PAM}}(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \overline{x}'_{ik})^2}{(s_i + s_0)^2} - 2 \cdot \log \pi_k,$$

where π_k is the prior probability of class k.

The classification rule is

$$C^{\mathrm{PAM}}(x^*) = l \text{ where } \delta_l^{\mathrm{PAM}}(x^*) = \min_k \{\delta_k^{\mathrm{PAM}}(x^*)\}$$

However, the *pamr* package [7] uses a different classifier. The *pamr* discriminant function for a sample x^* is

$$\delta_k^{\mathrm{PAMR}}(x^*) = \sum_{i=1}^{p} d'_{ik} \cdot m_k \cdot (\frac{x_i^* - \overline{x}_i}{s_0 + s_i} - \frac{1}{2} \cdot d'_{ik} \cdot m_k) + \log \pi_k$$

and the classification rule is

$$C^{\mathrm{PAMR}}(x^*) = l \text{ where } \delta_l^{\mathrm{PAMR}}(x^*) = \max_k \{\delta_k^{\mathrm{PAMR}}(x^*)\}$$

To distinguish between the classifier described in the paper and the classifier employed in *pamr*, we call the former the **PAM classifier** and the latter the **PAMR classifier**.

The Nearest Prototype (NP) Classifier (also called nearest mean classifier [9]) assigns a sample to the class whose prototype (one prototype per class) has the smallest Euclidian distance to the sample. The discriminant function for a sample x^* is: $\delta_k^{\mathrm{NP}}(x^*) = \|x^* - \overline{x}_k\|_2$, where $\overline{x}_k$ is the prototype of class k.

The classification rule is then: $C^{\mathrm{NP}}(x^*) = l$ where $\delta_l^{\mathrm{NP}}(x^*) = \min_k \{\delta_k^{\mathrm{NP}}(x^*)\}$.

Linear Discriminant Analysis (LDA) calculates a hyperplane in feature space. A quadratic discriminant function for a sample x^* is

$$\delta_k^{\mathrm{QDA}}(x^*) = (x^* - \overline{x}_k)^T \cdot S^{-1} \cdot (x^* - \overline{x}_k) - 2 \cdot \log \pi_k,$$

where $\overline{x}_k$ is a vector of centroids of class k in the features, S is the pooled estimate of the within-class covariance matrix and π_k is the prior probability of class k. This generally hyperquadratic decision surface is reduced to a hyperplane through our additional assumption of equal covariance matrices.

Support Vector Machines (SVMs) are learning machines that can be applied to classification problems. By solving a constrained quadratic optimization problem they can find a hyperplane that linearly separates a data set [10, 6]. Here, we use only the linear kernel.

2.3 Data Sets

The Golub Data Set contains leukemia microarray data with originally 6817 genes, 72 samples (47 ALL and 25 AML) and 2 classes. It was first analyzed by Golub et al. [11]. The preprocessing described by Dudoit et al. [12] was applied, reducing the number of features to 3051.

The Khan Data Set of small round blue cell tumors was analyzed by Khan et al. [13] and also by Tibshirani et al. in [1]. It consists of 63 samples (SRBCT classes: 23 EWS, 20 RMS, 12 NB, 8 NHL cases) with 2308 features in 4 classes.

The Diagnostic Chip Data Set of pancreatic tumors contains 62 samples (37 PaCa and 25 Pitis/Norm) with originally 558 features that were reduced to

169 genes and 2 classes [14]. The genes of this data set are known to be indicative for cancer diseases.

2.4 Testing Methods

Two different kinds of experiments were performed.

Predefined Feature Number. For comparing the feature reduction methods by the number of remaining features, the number of remaining features was fixed to values $p' = 5, 10, \ldots, 100$. This requires a special treatment of the PAM thresholds (see Sect. 2.1).

We tested each classifier separately on all feature reduction methods. SVMs were only applied on PCA feature reduction. With SVMs and ROC curves being suitable only for 2-class-problems (in the present configuration, i.e. we did not extend these methods to problem with more than two classes), we did not apply them to the Khan data set. In addition, the SVM and the NP classifiers were tested on the original (unreduced) feature sets.

The following testing methods were applied:

- Reclassification, that is, classification of the training data
- Leave-one-out, i. e. training the classifiers with all but one samples and classifying the remaining sample. This was repeated for all samples and the errors were added.
- 10×5-fold cross-validation: The samples were divided into 5 random groups. Each of the groups was left out once in training and used for classification. The errors were added. The procedure was repeated 10 times and the average error of the 10 runs was calculated.
- 10×10-fold cross-validation with 10 random groups.

Nested Cross-Vaildation for Feature Number Determination. For comparing the feature reduction methods by their optimal number of features, we applied a **nested 10-fold cross-validation**. This means a 10×5-fold cross-validation and a 10×10-fold cross-validation were performed, and the optimal number of features was chosen in each fold of this outer cross-validation by a nested 1×10-fold cross-validation on the training data of the current fold. This is the method proposed by Tibshirani et al. [1] for the PAM threshold selection. As this method cannot be applied reasonably on random methods, it was only performed on PAM, PCA and ROC feature reduction.

- For PAM, 30 sequential threshold values (the default thresholds chosen by the *pamr* package) were cross-validated and the threshold value that lead to the minimum number of errors at the first level and the minimum number of remaining features at the second level was used.
- For the other feature reduction methods, reductions to $p' = 5, 10, \ldots, 100$ features were cross-validated and the transformation that yielded the minimum number of errors at the minimum number of features was chosen.

3 Results

3.1 Experiments with Fixed Numbers of Features

The testing methods vary in their estimation of generalization error. While reclassification is usually an overly optimistic estimation of generalization error, different types of cross-validation give different hints regarding the match of classifier to data. We performed three types of cross-validation, namely leave-one-out, 10-fold and 5-fold cross-validation. For lack of space we show only the 5-fold cross-validation results in this section, they exhibit the highest error rates. Complete simulation results are given in the supplementary information available at http://www.informatik.uni-ulm.de/ni/mitarbeiter/HKestler/featred/.

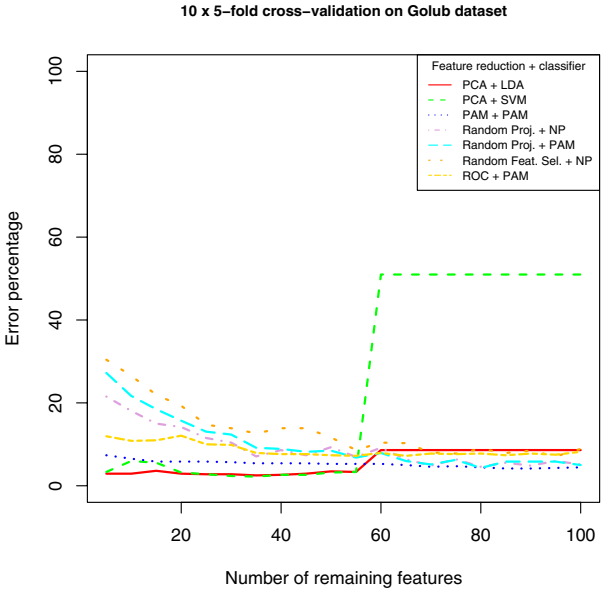
Golub Data Set

On the Golub data set all classifiers perform reasonably good with average error rates below 10/72 in a 5-fold cross-validation (see Figure 1). With the PAM and PAMR classifiers, the ROC feature reduction and PAM bring the best results in the cross-validations. In the 5-fold cross-validation, both PAM and PAMR start with 5.3 errors for 5 features and finally produce 3.0 errors with 80 features. None of the classifiers achieves an error of zero.

With the NP classifier, PCA is substantially better than other feature reduction methods for up to 55 features, yielding an average error of 1.8 cases or less in the 5-fold cross-validation. PCA error rates rise when the number of features increases. Starting with 70 features, Random Projection achieves comparable results to PAM with NP, leading to error rates mostly below 4.0/72. RP+NP even outperforms PAM+PAMR and PAM+PAM sometimes on feature numbers above 70. This result is surprising, but shows that Random Projection preserves distances quite well and works well with distance-based classifiers. With the LDA classifier, PCA feature reduction also brings good results, with error rates below 2.5/72 in the 5-fold cross-validation, while all other feature reduction methods produce even more errors with an increasing number of features. PCA and LDA also return constant zero error rates in reclassification, with 15 features and more. The SVM classifier yields the overall minimum error of 1.6/72 at 35 features with the PCA feature reduction in the 5-fold cross-validation, but error rates raise quickly with more features. Random Projection is always slightly better than Random Feature Selection and approaches or sometimes even beats PAM as the number of features rises.

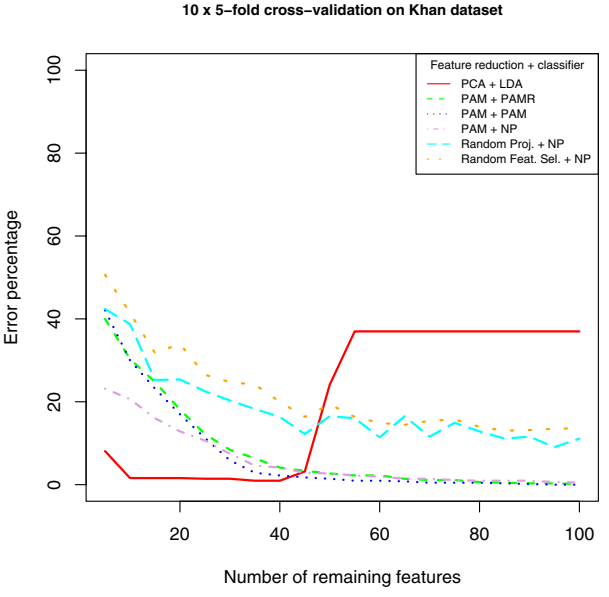
Khan Data Set

Compared to the Golub data, the Khan data set is harder to classify and needs more features for correct classification. With a low number of features, PCA always performs best on Khan (see Figure 2). Here again, it shows that LDA and PCA go together well, leading to an average of 1 error in the 5-fold cross-validation with only 15 features. With more than 40 features, the PAM feature reduction performs better than PCA, which gets worse. Yet, PAM achieves less



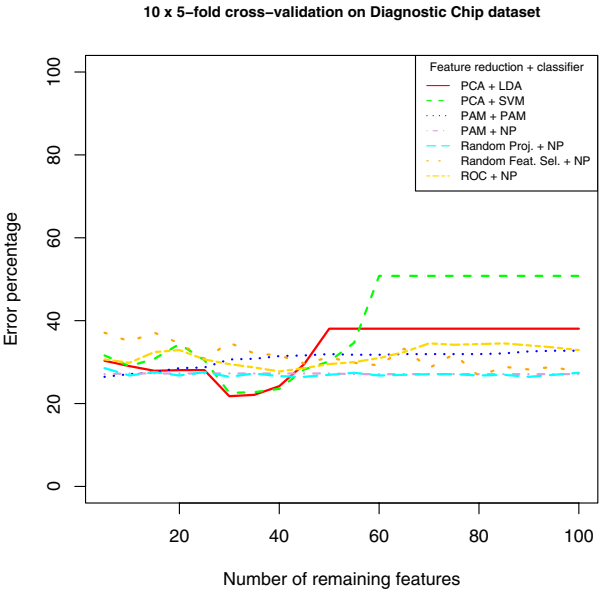
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
PCA + LDA	2.9	2.9	3.6	2.9	2.8	2.8	2.5	2.6	2.9	3.5	3.3	8.6	8.6	8.6	8.6	8.6	8.6	8.6	8.6	8.6
PCA + SVM	3.3	6.0	5.6	3.2	2.8	2.4	2.2	2.6	2.6	3.2	3.3	51.0	51.0	51.0	51.0	51.0	51.0	51.0	51.0	51.0
PAM + PAM	7.4	6.5	5.8	5.8	5.8	5.7	5.4	5.4	5.4	5.3	5.3	5.3	5.0	4.6	4.7	4.4	4.2	4.2	4.3	4.4
Random Proj. + NP	21.5	18.1	15.0	14.2	11.5	10.4	7.1	8.6	7.4	9.3	6.9	9.2	6.4	4.9	6.4	4.6	5.6	4.9	6.0	5.1
Random Proj. + PAM	27.2	21.7	18.5	15.7	13.1	12.4	9.2	8.9	8.2	8.5	6.8	7.9	6.0	5.1	6.3	4.3	5.8	5.8	5.8	5.0
Random Feat. Sel. + NP	30.4	26.5	21.9	19.3	14.7	13.9	12.6	13.9	13.9	11.7	8.6	10.4	10.3	8.1	7.8	8.9	7.9	8.5	7.1	8.9
ROC + PAM	11.9	10.8	11.0	12.1	10.0	9.9	7.9	7.6	7.6	7.4	7.2	7.9	7.2	7.8	7.6	7.8	7.4	7.8	7.5	8.2

Fig. 1. Golub data set: Error rates of the five-fold cross-validation. Random Projection performs surprisingly well with higher numbers of features. Shown are the error rates for the combinations: ROC+PAM, RF+NP, RP+PAM, RP+NP, PAM+PAM, PCA+SVM, PCA+LDA.



	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
PCA + LDA	8.1	1.6	1.6	1.6	1.4	1.4	1.0	1.0	3.2	24.1	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0
PAM + PAMR	40.0	30.3	24.8	18.1	12.2	8.4	6.3	4.1	3.3	2.7	2.2	2.2	1.4	1.1	1.1	0.6	0.5	0.3	0.3	0.2
PAM + PAM	42.1	30.0	23.2	17.0	11.3	5.9	2.9	2.2	1.7	1.4	1.0	1.0	0.8	0.5	0.5	0.5	0.3	0.2	0.0	0.0
PAM + NP	23.2	20.6	16.0	12.9	10.6	7.5	4.6	4.1	2.9	2.7	2.2	1.9	1.4	1.4	1.1	1.0	1.0	1.0	0.6	0.5
Random Proj. + NP	42.4	38.7	25.2	25.4	22.5	20.3	18.3	16.3	12.2	16.5	16.0	11.4	16.5	11.6	14.9	12.9	11.1	11.6	9.0	11.1
Random Feat. Sel. + NP	50.6	41.4	31.7	34.0	26.5	24.8	24.1	20.0	16.3	19.4	16.3	14.9	14.4	15.4	15.7	14.0	13.0	13.2	13.5	13.7

Fig. 2. Khan data set: Error rates of the five-fold cross-validation. PCA and LDA perform well with a small number of features, NP outperforms the PAM classifiers with less than 30 features. Shown are the error rates for the combinations: RF+NP, RP+NP, PAM+NP, PAM+PAM, PAM+PAMR, PCA+LDA.



	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
PCA + LDA	30.3	29.0	27.9	28.1	28.1	21.8	22.1	24.2	29.5	38.1	38.1	38.1	38.1	38.1	38.1	38.1	38.1	38.1	38.1	38.1
PCA + SVM	31.6	29.0	30.8	34.4	30.5	22.6	22.7	23.5	28.2	30.2	34.7	50.8	50.8	50.8	50.8	50.8	50.8	50.8	50.8	50.8
PAM + PAM	26.5	27.1	27.6	28.5	28.7	30.6	30.8	31.5	31.6	31.9	31.8	31.8	31.9	31.9	31.9	31.9	32.1	32.6	32.7	32.7
PAM + NP	27.1	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.3	27.1	27.1	27.1	27.1	27.1	27.1	27.1	27.1	27.1	27.1
Random Proj. + NP	28.5	26.8	27.6	26.8	27.6	26.5	27.3	26.6	26.5	26.9	27.4	26.8	26.9	27.1	27.1	26.8	26.9	26.5	26.9	27.4
Random Feat. Sel. + NP	37.1	35.0	37.1	34.5	30.8	34.7	32.1	31.6	29.7	31.1	29.8	29.2	33.2	28.5	31.9	26.6	28.9	28.2	28.9	27.4
ROC + NP	30.6	29.8	32.4	32.9	30.6	29.5	28.7	27.7	28.5	29.5	30.0	31.0	32.4	34.5	34.2	34.4	34.5	34.0	33.5	32.9

Fig. 3. Diagnostic Chip data: Error rates of the five-fold cross-validation. No feature reduction and classifier combination produces acceptable results. Shown are the error rates for the combinations: ROC+PAM, RF+NP, RP+PAM, RP+NP, PAM+PAM, PCA+SVM, PCA+LDA.

than 1 error at 65 features with the PAMR classifier and at 50 features with the PAM classifier. The PAMR and PAM classifiers show similar behaviours, with PAM being slightly better than PAMR in most cases. NP shows analogue characteristics, but yields error rates below those of PAMR and PAM with less than 30 features. With more than 30 features, PAMR and PAM perform slightly better. As with the Golub data set, Random Projection always performs better than Random Feature Selection. However, it does not reach the PAM classifier with the Khan data set.

Diagnostic Chip Data Set

All classifiers achieve the highest error rates on the Diagnostic Chip data set. None of the classifiers achieves an error below 10 (SVM: error rate 10.9/62 on the complete data, no feature reduction, 10×5-fold cv). The best result with feature reduction is achieved by PCA and LDA classification with a minimum average error of 13.5 cases with 30 features (10×5-fold cv) and PCA and SVM with a minimum average error of 14 cases with 30 features (10×5-fold cv), see Figure 3. With the PAM and PAMR classifiers, PAM feature reduction leads to bad results, sometimes even worse than Random Feature Selection. Both classifiers produce exactly the same error rates in the 5-fold cross-validation. PCA, Random Projection and ROC are better. With increasing numbers of features, PCA performs best with an error rate of 16.5 in the 5-fold cross-validation starting with 50 features.

The Nearest Prototype classifier returns almost constant error rates of around 17 errors (min 15) for all numbers of features when being applied on feature sets produced by PCA, Random Projection and PAM. The error rates of the other feature reduction methods with the NP classifier are unstable. Without feature reduction, the average error is 22.6 cases. In contrast to the other data sets, the error rates of Random Projection in the different runs are quite stable, i. e. have a low variance. Those facts indicate that it is hard to find a subset of features that suits significantly better for distance-based classification than the original feature set and that none of the feature reduction methods succeeds in finding one. Consequently, the PAM and the PAMR classifiers, which include distance measurements as well, also mostly return error rates that do not change very much with different numbers of features. Only the PAM feature reduction and ROC show slightly increasing error rates with an increasing number of features.

3.2 Experiments with Nested Cross-Validations

Golub Data Set

PAM seems not to be able to find a stable number of features on Golub. In the 5-fold cross-validation, it determines an average of 1243.9 features for the minimum error of 3.1/72, with a very high standard deviation of 322.3 features. 10-fold cross-validation produces a similar deviation, but a mean value of 1601.0

features and an error of 3.4/72. In fact, PAM achieved even better results at a much lower number of features in the experiments with fixed numbers of features, i. e., 3.0/72 errors at 85 features in the 5-fold cross-validation and 3.0/72 errors at 90 features in the 10-fold cross-validation. This may be due to the fact that the 30 pre-defined thresholds used by the original PAM cover the whole feature set, while our 20 thresholds in the above experiments lead to 5-100 features and thus have a smaller raster in this area. With PCA, the results are mostly comparable to our previous experiments. LDA achieves a mean error of 3.1/72 at an average of 10.9 features in the 5-fold cross-validation. The best result is achieved by the linear SVM at 16.2 features with an error of 2.7/72. ROC feature reduction only produces average errors of 4.9/72 (linear SVM) and more in the 5-fold cross-validation.

Khan Data Set

On Khan, PAM produces more stable results, yielding a mean error of 1.4/63 at an average of 52.7 features in the 5-fold cross-validation. Anyway, in the fixed-feature experiments, the error rate grew smaller continuously towards 100 features, so this is still not an optimal value. The standard deviation of the number of features is 7.8, while the PCA standard deviation is less than 2.8, depending on the classifier. PCA again performs well with LDA, producing an error of 1.5/63 at only 14.8 features. The results of the fixed-feature experiments were slightly better. The NP classifier achieves an error of 3.6 at 16.1 features in the 5-fold cross-validation, which is also comparable to the previous experiments.

Diagnostic Chip Data Set

As already seen in the fixed-feature experiments, the error rates on this data set are high. PAM achieves 17/62 errors at 5.7 features (5-fold cross-validation), which is slightly worse than the previous results (16.4 errors at 5 features). The number of features is comparatively stable with a deviation of 1.4. With PCA, NP achieves similar results with 16.9/62 errors at an average of 5.4 features with a very small standard deviation of 0.84. This time, LDA only yields 18.5/62 errors at 22 features with a standard deviation of 5.1. The error rate is comparable with the error rate of the fixed-feature experiments, but the number of features chosen is not the one that produced a minimum error there. With ROC feature reduction, LDA yields a similar error (18.8/62), but needs only 13 features with a smaller deviation of 1.5. The SVM produces a minimum error of 16.4 with 26.1 features. NP performs worse with an error of 19.9/62 at 26 features with a high standard deviation of 7.9 features.

4 Discussion

The experimental results show that especially when classifying with small numbers of features, PAM is not the best choice. In particular, PCA feature reduction

in combination with the LDA and SVM classifiers performs excellently with very low-dimensional target feature spaces.

In addition, the optimal number of features returned by PCA in the nested cross-validations was quite stable, indicating that the feature selection remains reliable when changing a few elements in the training set.

However, the factors returned by PCA cannot be interpreted easily in order to determine which features in the original feature space are important for classification. Obviously, both the LDA classifier and the SVM classifier (linear kernel) lead to high classification errors when over-fitted, hence they seem not suitable for a larger number of features. Also the different types of microarrays, e.g. whole genome array vs selected genes, seem to play a major role in the classifier performance (together with the type of tissue). For the Golub and Khan data sets genes (n= 6817) were not specifically selected for discrimination, whereas the genes (n= 558) for the pancreas vs pancreatitis diagnostic chip data were selected according to their known or believed involvement in cancer.

Generally speaking, PAM does not stand out compared to other feature reduction methods. The only data set showing major distances between the error rates of PAM and the baseline algorithms Random Feature Selection and Random Projection is the Khan data set, where there are differences of 10 to 15 in the cross-validations. In the other data sets, the random methods are mostly slightly worse, but sometimes even outperform PAM. The nested cross-validation experiments mostly did not find the number of features that lead to the minimum error in the previous fixed-feature experiments which in itself is not surprising. This may be due to the choice of the feasible threshold values in PAM leading sometimes to zero remaining features or including all features in the decision process which is not desirable (linear separability). In contrast, using all possible thresholds that lead to different feature numbers (in a certain range) might avoid this.

Random Projection was observed to perform better than Random Feature Selection. The good results of Random Projection in combination with the Nearest Prototype classifier underline its capability to preserve distances and show that Random Projection is used most effectively with distance-based classifiers. A possible explanation for this phenomenon is the fact that Random Feature Selection completely discards features by random, while Random Projection projects all features to the lower-dimensional space. Thus, Random Projection may not lose as much information as Random Feature Selection. This may shed a new light on an often used premise which is used in normalization procedures, that only a small fraction of genes is regulated in a gene expression microarray experiment.

Acknowledgments

We thank André Müller for fruitful discussions. This work is supported by the Stifterverband für die Deutsche Wissenschaft (HAK) and the German Science Foundation, SFB 518, Project C5.

References

1. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS **99** (2002) 6567–6572
2. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. Statistical Science **18** (2003) 104 – 117
3. Vempala, S.: The Random Projection Method. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, volume 65 (2004)
4. Duda, R.O., Hart, P., Storck, D.: Pattern classification. Wiley (2001)
5. Webb, A.: Statistical Pattern Recognition. Wiley (2002)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press (2000)
7. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: PAMR package version 1.27, http://www-stat.stanford.edu/~tibs/PAM/Rdist (2005)
8. Johnson, W., Lindenstrauss, J.: Extensions of Lipshitz mapping into Hilbert space. Contemporary Mathematics **26** (1984) 189–206
9. Therrien, C.: Decision estimation and classification. Wiley (1989)
10. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In Haussler, D., ed.: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press (1992) 144–152
11. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531 – 536
12. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association **97** (2002) 77 –87
13. Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., Meltzer, P.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine **7** (2001) 673 – 679
14. Buchholz, M., Kestler, H., Bauer, A., Bock, W., Rau, B., Leder, G., Kratzer, W., Bommer, M., Scarpa, A., Schilling, M., Adler, G., Hoheisel, J., Gress, T.: Specialized DNA arrays for the differentiation of pancreatic tumors. Clin. Cancer Res. **11** (2005) 8048–54