Linear Contrast Classifiers in High-Dimensional Spaces

Florian Schmid*, Ludwig Lausser*, and Hans A. Kestler**

Medical Systems Biology and
Institute of Neural Information Processing
Ulm University, 89069 Ulm, Germany
`hans.kestler@uni-ulm.de`

Abstract. Linear classifiers are mainly discussed in terms of training algorithms that try to find an optimal hyperplane according to a data dependent objective. Such objectives might be the induction of a large margin or the reduction of the number of involved features. The underlying concept class of linear classifiers is analyzed less frequently. It is implicitly assumed that all classifiers of this function class share the same common properties.

In this work we analyze the concept class of linear classifiers. We show that it includes different subclasses that show beneficial properties during the prediction phase. These properties can directly be derived from the structural form of the classifiers and must not be learned in a data dependent training phase.

We describe the concept class of contrasts, a class of linear functions that is for example utilized in variance analyses. Models from this concept class share the common property of being invariant against global additive effects. We give a theoretical characterization of contrast classifiers and analyze the effects of replacing general linear classifiers by these new models in standard training algorithms.

1 Introduction

The limited availability of samples is one of the major challenges of the development of diagnostic models for biomolecular data. Especially for high-dimensional gene expression profiles as obtained from microarray experiments the low cardinality of a dataset can lead to non generalizable models. Possible reasons for small sample sizes may be the rareness of a disease, high costs or technical burdens. A possible strategy for enlarging a dataset can be to merge the samples of different research groups. These collections might be affected by an additional source of variation. Being geographically separated the experimental conditions may not be identical in different labs. Environmental factors such as humidity and temperature are likely to vary for different locations. They can affect all simultaneous measurements of a sample (e.g. all measurements of a gene expression profile). These effects should be minimized during the development of

* F. Schmid and L. Lausser contributed equally.

** Corresponding author.

a classification model. In this work we propose a linear classification model that may be suitable for this scenario.

The research on linear classifiers mainly focuses on the development of training algorithms and optimization criteria [7]. Famous example are linear support vector machines that maximize the margin between training samples [14] or the nearest shrunken centroid classifier that utilizes embedded feature selection [13]. The structural properties of linear classifiers are often neglected. It is implicitly assumed that all classifiers of this function class share the same common properties.

We recently showed that subclasses of linear classifiers exist which share common structural properties that induce an invariance against a certain type of noise [8]. That is a classifier is guaranteed to be unaffected by this type of data transformation, if chosen from this subclass. In this work we present an additional subclass, which we call linear contrast classifiers. It consists of linear classifiers that are invariant against global transition. In the following we will prove the invariance property of this subclass. We will also characterize the benefit of contrast classifiers in different noise experiments.

2 Methods

In this work a classifier c will be seen as a function mapping that predicts the class label $y \in \mathcal{Y}$ of an object according to a vector of measurements $\mathbf{x} \in \mathcal{X}$. The components of a vector will be denoted as $\mathbf{x} = (x^{(1)}, \dots, x^{(n)})^T$. In the following we restrict ourselves to binary classification tasks (e.g. $\mathcal{Y} = \{0, 1\}$) and $\mathcal{X} \subseteq \mathbb{R}^n$. A classifier will be selected out of a concept class of functions sharing some structural property (e.g. linear classifiers) $c \in \mathcal{C}$. The selection or training of a classifier is normally done according to a performance measure evaluated on a set of labeled training examples $\mathcal{S}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. The generalization ability of the selected classifier is estimated on an independent set of test samples $\mathcal{S}_{te} = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{m'}$.

In the following the concept class of linear classifiers $\mathcal{C}_{lin}$ is analyzed.

$$\mathcal{C}_{lin} = \left\{ \mathbb{I}_{[\langle \mathbf{w}, \mathbf{x} \rangle \geq t]} \mid \mathbf{w} \in \mathbb{R}^n, t \in \mathbb{R} \right\} \quad (1)$$

A classifier $c \in \mathcal{C}_{lin}$ can be interpreted as a linear hyperplane separating the samples into two classes. It is typically parameterized by a weight vector $\mathbf{w} \in \mathbb{R}^n$ determining its orientation and a threshold or offset $t \in \mathbb{R}$ determining its distance to the origin (Figure 1).

2.1 Invariant Classification Models

Reducing the complexity of a concept class is not the only reason for searching for structural subgroups. Another reason might be the gain of invariances that allow to completely neglect a certain kind of influence on the label prediction.

Fig. 1. Scheme of a linear classifier: The decision boundary of a linear classifier is shown. The shade of the samples indicates the corresponding classification. Samples in light gray are categorized as class 1 and samples in dark gray as class 0.

Definition 1. A classifier $c : \mathcal{X} \rightarrow \mathcal{Y}$ is called invariant against a parameterized class of data transformations $f_\theta : \mathcal{X} \rightarrow \mathcal{X}$ if

$$\forall \theta, \forall \mathbf{x} \in \mathcal{X} : c(f_\theta(\mathbf{x})) = c(\mathbf{x}). \quad (2)$$

A concept class $\mathcal{C}$ is called invariant against f_θ if each $c \in \mathcal{C}$ is invariant against f_θ .

Choosing a classifier out of an invariant concept class ensures that a classification result is completely unaffected by a misleading transformation $f_\theta : \mathcal{X} \rightarrow \mathcal{X}$. Invariant classifiers are especially beneficial if the measurements of an object are known to be distorted by a technical process.

An example of an invariant concept class is the concept class of offset-free linear classifiers $\mathcal{C}_0$.

$$\mathcal{C}_0 = \{ \mathbb{I}_{[\langle \mathbf{w}, \mathbf{x} \rangle \geq 0]} \mid \mathbf{w} \in \mathbb{R}^n \} \quad (3)$$

These classifiers omit the adaption of an offset ($t = 0$) [8]. Offset-free linear classifiers are invariant against the global scaling of the data $f_a(\mathbf{x}) \mapsto a\mathbf{x}$, $a \in \mathbb{R}_+$.

The invariance of a concept class is also of interest for classifier ensembles $\mathcal{E}$

$$\forall c \in \mathcal{E} : \mathcal{X} \rightarrow \mathcal{Y}, \quad h_{\mathcal{E}} : \mathcal{Y}^{|\mathcal{E}|} \rightarrow \mathcal{Y} \quad (4)$$

More precisely Equation 4 describes a fusion strategy $h_{\mathcal{E}}$ that operates on the predicted class labels of the base classifiers $c \in \mathcal{E}$ (late aggregation) [12]. Choosing all base classifiers from an invariant concept class $\mathcal{E} \subset \mathcal{C}$ the fusion strategy gets itself invariant against f_θ .

$$h_{\mathcal{E}}(c_1(f_{\theta_1}(\mathbf{x})), \dots, c_{|\mathcal{E}|}(f_{\theta_{|\mathcal{E}|}}(\mathbf{x}))) = h_{\mathcal{E}}(c_1(\mathbf{x}), \dots, c_k(\mathbf{x})) \quad (5)$$

If the base classifiers operate on distinct feature spaces, the invariance property is able to counteract local effects affecting only subgroups of measurements.

2.2 Contrast Classifiers

In this work we study an invariant subclass of linear classifiers $\mathcal{C}_{con} \subset \mathcal{C}_{lin}$

$$\mathcal{C}_{con} = \left\{ \mathbb{I}_{[\langle \mathbf{w}, \mathbf{x} \rangle \geq t]} \left| \sum_{i=1}^n w^{(i)} = 0, \mathbf{w} \in \mathbb{R}^n, t \in \mathbb{R} \right. \right\}. \quad (6)$$

For these linear classifiers, the orientation of a hyperplane $\mathbf{w}$ is additionally constraint by $\sum_{i=1}^n w^{(i)} = 0$. The choice of t is equal for $\mathcal{C}_{lin}$ and $\mathcal{C}_{con}$. Linear functions that fulfill this property are called *contrasts* in variation analysis [5]. We therefore call $\mathcal{C}_{con}$ the concept class of (linear) contrast classifiers. They can be shown to be invariant against global translation.

Theorem 1. *The concept class of contrast classifiers $\mathcal{C}_{con}$ as defined in Equation 6 is invariant against global translation. That is*

$$\forall c \in \mathcal{C}_{con}, \forall b \in \mathbb{R}, \forall \mathbf{x} \in \mathcal{X} : \quad c(f_b(\mathbf{x})) = c(\mathbf{x}), \quad (7)$$

where $f_b \mapsto \mathbf{x} + \mathbf{b}$ and $\mathbf{b} = (b, \dots, b)^T$, $\mathbf{b} \in \mathbb{R}^n$.

Proof (Theorem 1). For proofing Theorem 1, it is sufficient to show that global translation does not affect the projection of a contrast.

$$\langle \mathbf{w}, \mathbf{x} + \mathbf{b} \rangle = \langle \mathbf{w}, \mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{b} \rangle \quad (8)$$

$$= \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^n w^{(i)} b \quad (9)$$

$$= \langle \mathbf{w}, \mathbf{x} \rangle + b \underbrace{\sum_{i=1}^n w^{(i)}}_{=0} \quad (10)$$

$$= \langle \mathbf{w}, \mathbf{x} \rangle \quad (11)$$

$$(12)$$

□

Figure 2 gives a two-dimensional example of a contrast classifier. In this low dimensional space contrast classifiers are restricted to two possible orientations $w \in \{(-1, +1)^T, (+1, -1)^T\}$. Only the threshold t can be modified. The orientation of a contrast classifiers becomes more flexible in higher dimensions. Figure 2a shows the invariance property of a contrast classifier. If a sample is moved by a global transition, the sample is guarantied to receive the same class label as before. Interestingly there exist some linear separable cases that can not be separated by a linear contrast classifier (Figure 2b).

Fig. 2. Scheme of a linear contrast classifier: In a two-dimensional space a contrast can only be adapted by modifying its threshold t (Panel a). The contrast classifier is invariant against global transitions of type $(x^{(1)}, x^{(2)})^T \pm (b, b)^T$. Panel b indicates the limitations of a linear contrast classifier. It shows an example of a linear separable dataset that can not be separated by a linear contrast classifier in two dimensions.

3 Experiments

In order to investigate the properties of contrast classifiers we have conducted experiments on artificial datasets (Section 3.2) and real datasets (Section 3.3). In general all these experiments are performed with help of the TunePareto R package [9].

3.1 Support Vector Machines

The influence of the contrast constraint is investigated in the context of training linear support vector machines (SVM) [14]. Linear SVMs are linear classifiers that optimize following objective:

$$\min_{\mathbf{w}, t, \xi} \quad \frac{1}{r} \|\mathbf{w}\|_r^r + C \sum_{i=1}^n \xi_i^l \quad (13)$$

$$\text{s.t.} \quad \forall i : y_i (\mathbf{w}^T \mathbf{x}_i - t) \geq 1 - \xi_i \quad (14)$$

$$\forall i : \xi_i \geq 0 \quad (15)$$

$$\sum_{i=1}^n w^{(i)} = 0 \text{ (only contrasts)} \quad (16)$$

The objective mainly consists of two terms (Equation 13). The first one is a regularization term based on a norm of $\mathbf{w}$. The second one is a loss term that

summarizes the losses $\xi_i \in \mathbb{R}_+$ of the constraints in Equation 14. Both terms are combined via a tradeoff parameter $C \in \mathbb{R}_+$. The training algorithm for linear SVMs can be modified to a training algorithm for linear contrasts by adding the constraint of Equation 16.

We consider four different types of SVMs [1], which we denote by $L1R2$ ($l = 1, r = 2$), $L2R2$ ($l = 2, r = 2$), $L1R1$ ($l = 1, r = 1$), $L2R1$ ($l = 2, r = 1$). The l parameter influences the effects of a single loss variable ξ_i . In contrast to a linear weighting ($l=1$), a value of $l = 2$ increases the influence of large values ($\xi_i > 1$) and decreases the influence of small values ($\xi_i < 1$). The r parameter defines the norm utilized for regularization. A value of $r = 2$ corresponds to the original support vector machine [14]. In this case the SVM training algorithm constructs a large margin classifier that maximizes the distance between the training samples and the hyperplane. A value of $r = 1$ leads to an embedded feature selection [2]. Smaller weights will directly be set to 0. In order to distinguish between SVMs based on linear classifiers and linear contrasts, we will mark the linear contrasts with the subscript *con* (e.g. $L2R1_{con}$).

3.2 Experiments on Artificial Data

We compare the new contrast classifiers to the standard support vector machines in sequences of experiments on artificial data. The datasets are generated by drawing samples from two normal distributions $\mathcal{N}(\mathbf{c}_y, \mathbf{I})$, $y \in \mathcal{Y}$. Here, $\mathbf{c}_y \in \mathbb{R}^n$ denotes the class wise centroids and $\mathbf{I} \in \mathbb{R}^{n \times n}$ denotes the identity matrix. The first centroid is chosen according to $\mathbf{c}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. The second one is calculated by $\mathbf{c}_0 = \mathbf{c}_1 + d\mathbf{w}/\|\mathbf{w}\|_2$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. In this way, it is guaranteed that $\|\mathbf{c}_1 - \mathbf{c}_0\|_2 = d$.

We generated different training/test scenarios by varying the dimensionality of the datasets $n \in \{2, 5, 10, 250, 500, 1000\}$ and the distance between class centroids $d \in \{1, 1.1, \dots, 5\}$. Each experiment is based on a training set of 50 samples per class. An independent test set of 2×50 samples is generated for evaluating the classifier. Each combination of n and d is tested for 5 different pairs of centroids.

3.3 Experiments on Real Data

The classifiers are also compared in experiments on real microarray datasets. The utilized datasets are given in Table 1. In these experiments, the classifiers are evaluated in an $r \times f$ cross validation [3]. That is, the data is split into f folds of approximately equal size. One of the folds is removed from the dataset before a classifier is trained. It is used for testing the classifier afterwards. The procedure is repeated for each combination of folds. The cross validation is repeated on r permutations of the original dataset. The error of the classifier is estimated by

$$R_{cv} = \frac{1}{rm} \sum_{i=1}^r \sum_{j=1}^f \sum_{(\mathbf{x}, y) \in \mathcal{S}_{te}^{i,j}} \mathbb{I} \left[c_{S_{tr}^{i,j}}(\mathbf{x}) \neq y \right]. \quad (17)$$

Here $\mathcal{S}_{tr}^{i,j}$ and $\mathcal{S}_{te}^{i,j}$ denote the training set and test set of the i th run and the j th split. We have chosen to set $r = 10$ and $f = 10$.

Besides the noise free experiment, we have also conducted experiments with two different types of artificial noise.

Noisy Test Samples: In this setting the samples of a test set $\mathcal{S}_{te} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m'}$ are affected by a random global transition $f_b \mapsto \mathbf{x} + \mathbf{b}$.

$$\mathcal{S}'_{te} = \{(f_{b_i}(\mathbf{x}_i), y_i)\}_{i=1}^{m'}, \quad \forall i : b_i \sim \mathcal{U}(-p, p) \quad (18)$$

Here $\mathcal{U}$ denotes a uniform distribution. The upper and lower limit of b_i is chosen according to $p \in \{1, \dots, 5\}$.

Noisy Training Samples: Here we assume that the training set $\mathcal{S}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is affected by a global transition. We model an offset of the samples of class 1.

$$\mathcal{S}'_{tr} = \{(f_{b_i}(\mathbf{x}_i), y_i)\}_{i=1}^m, \quad b_i = \begin{cases} p & \text{if } y_i = 1 \\ 0 & \text{else} \end{cases} \quad (19)$$

Table 1. The table summarizes the four analyzed microarray datasets. The number of features n , the class-wise numbers of samples m_y , $y \in \{0, 1\}$ and the analyzed phenotypes are shown.

name	n	m_1	m_0	phenotypes
Buchholz [4]	169	37	25	pancreatic adenocarcinoma vs pancreatitis subclasses of bladder cancer
Dyrskj�t [6]	7071	20	20	
Pomeroy [10]	7071	9	25	classic vs desmoplastic medulloblastoma
Shipp [11]	7071	19	58	diffuse large B-cell lymphoma vs follicular lymphoma

4 Results

The results of the experiments on artificial data are shown in Figure 3. It can be seen that the standard linear classifiers perform better than their contrast equivalents in low dimensional spaces ($n \in \{2, 5\}$). If the dimensionality is increased it can be seen that the results of the contrasts get closer to those of the standard linear classifiers ($n \in \{10, 250\}$). For the datasets with highest dimensionality ($n \in \{500, 1000\}$) no difference between the standard linear classifiers and their corresponding contrasts can be observed.

For each combination of dimensionality and pair of standard linear classifier and corresponding contrast a two-sided Wilcoxon test has been performed. The resulting p-values are all greater 0.1308 for $n \in \{250, 500, 1000\}$.

Fig. 3. Results of the artificial experiments. Each row shows the mean differences in accuracy between a support vector machine ($L1R1$, $L1R2$, $L2R1$, $L2R2$) and the corresponding contrast classifier ($L1R1_{con}$, $L1R2_{con}$, $L2R1_{con}$, $L2R2_{con}$). The single columns show experiments with different dimensionality $n \in \{2, 5, 10, 250, 500, 1000\}$. In each cell, the distance between the class centroids is varied $d \in \{1.0, 1.1, \dots, 5.0\}$. A value above 0 indicates a better performance of the standard support vector machine. A value below 0 indicates a better performance of the contrast classifier.

Table 2 shows the 10×10 cross-validation errors of the noise free experiments on the real datasets. For each of the datasets, one of the contrast classifiers outperforms all other models. Over all datasets the $L2R1$ support vector machines perform better or equal than all other tested classifiers. Interestingly in most of the cases the contrast classifiers outperform the standard linear classifiers. Exceptions for this are the $L1R1_{con}$ on the Buchholz and the Pomeroy dataset.

The results for the experiments with noisy test samples are given in Figure 4. While the contrast classifiers are not affected by the added noise the performance of the standard linear classifiers varies. In most of the cases the error rates are increased or equal compared to the zero noise cases. Only in a few combinations the error rate is slightly decreased for the standard linear classifiers.

The results for noisy training samples are shown in Figure 5. It can be seen that the results of the standard linear classifiers (except the $L1R1$ on the Pomeroy dataset) become worse by an increasing noise level. The contrast classifiers perform equally across all experiments and are not affected by noise.

Fig. 4. Noisy test sample experiment on real data: The 10×10 cross-validation errors for different levels of noise are shown. Each cell includes the results for a combination of dataset and pair of classifiers (standard vs corresponding contrast classifier).

Table 2. Results of the noise free 10×10 cross validation experiments on the real datasets. The mean error rates are shown. Lowest errors per dataset are given in bold.

	Buchholz [4]	Dyrskjøt [6]	Pomeroy [10]	Shipp [11]
<i>L1R1</i>	25.16	10.25	24.70	4.94
<i>L1R1_{con}</i>	27.58	9.00	25.59	4.03
<i>L1R2</i>	16.45	4.50	15.59	2.08
<i>L1R2_{con}</i>	16.29	4.50	15.59	2.08
<i>L2R1</i>	15.65	4.50	15.59	2.08
<i>L2R1_{con}</i>	15.32	4.50	15.59	1.95
<i>L2R2</i>	16.45	4.50	15.59	2.08
<i>L2R2_{con}</i>	16.29	4.50	15.59	2.08

5 Conclusion

Contrast classifiers describe a functional subclass of linear classifiers that are invariant against global transition. This property can directly be related to the

Fig. 5. Noisy training sample experiment on real data: The 10×10 cross-validation errors for different levels of noise are shown. Each cell includes the results for a combination of dataset and pair of classifiers (standard vs corresponding contrast classifier).

structure of a classifier and is independent from a particular training process or a particular dataset.

Our experiments show that invariances can be beneficial, if the samples can not be guaranteed to be affected by a unique source of noise. This is at least the case, if a classifier is trained under well defined conditions and afterwards applied in a real life scenario on noisy test samples. Invariances might also be of interest for clinical multi center studies, which are based on data collected in different laboratories. Here, global effects can also affect the training process of a classifier. Our experiments suggest that biased training samples can be even more corruptive for the prediction accuracy on further samples than noisy test samples.

Nevertheless, an invariance property also reduces the flexibility of a classifier and therefore decreases its complexity. In the case of contrast classifiers, our experiments show that this is only an limitation in low-dimensional settings. In high-dimensional spaces the accuracies of linear classifiers with and without invariance properties are comparable. For lower dimensional settings it may be

an option to develop classifier ensembles of contrast classifiers as they inherit the invariance property of the base learners.

Acknowledgements. The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007–2013) under grant agreement n°602783 to HAK, the German Research Foundation (DFG, SFB 1074 project Z1 to HAK), and the Federal Ministry of Education and Research (BMBF, Gerontosys II, Forschungskern SyStaR, project ID 0315894A to HAK).

References

1. Abe, S.: Support Vector Machines for Pattern Classification. Springer (2010)
2. Bhattacharyya, C., Grate, L., Rizki, A., Radisky, D., Molina, F., Jordan, M., Bissell, M., Mian, I.: Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data. *Signal Processing* 83(4), 729–743 (2003)
3. Bishop, C.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer (2006)
4. Buchholz, M., Kestler, H.A., Bauer, A., Böck, W., Rau, B., Leder, G., Kratzer, W., Bommer, M., Scarpa, A., Schilling, M.K., Adler, G., Hoheisel, J.D., Gress, T.M.: Specialized DNA arrays for the differentiation of pancreatic tumors. *Clinical Cancer Research* 11(22), 8048–8054 (2005)
5. Casella, G., Berger, R.: Statistical Inference. Thomson Learning (2002)
6. Dyrskjöt, L., Thykjaer, T., Kruhøffer, M., Jensen, J., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., Orntoft, T.: Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics* 33(1), 90–96 (2003)
7. Lausser, L., Kestler, H.A.: Robustness analysis of eleven linear classifiers in extremely high-dimensional feature spaces. In: Schwenker, F., El Gayar, N. (eds.) *ANNPR 2010, LNCS*, vol. 5998, pp. 72–83. Springer, Heidelberg (2010)
8. Lausser, L., Kestler, H.: Fold change classifiers for the analysis of gene expression profiles. In: Gaul, W., Geyer-Schulz, A., Baba, Y., Okada, A. (eds.) *Proceedings volume of the German/Japanese Workshops in 2010 (Karlsruhe) and 2012 (Kyoto), Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 193–202 (2014)
9. Müssel, C., Lausser, L., Maucher, M., Kestler, H.: Multi-objective parameter selection for classifiers. *Journal of Statistical Software* 46(5), 1–27 (2012)
10. Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., Golub, T.: Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415(6870), 436–442 (2002)
11. Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T., Mesirov, J., Neuberg, D., Lander, E., Aster, J., Golub, T.: Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8(1), 68–74 (2002)

12. Tulyakov, S., Jaeger, S., Govindaraju, V., Doermann, D.: Review of classifier combination methods. In: Marinai, S., Fujisawa, H. (eds.) *Machine Learning in Document Analysis and Recognition*, pp. 361–386. Springer (2008)
13. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* 99(10), 6567–6572 (2002)
14. Vapnik, V.: *Statistical Learning Theory*. Wiley (1998)