Classification and Feature Selection on Matrix Data with Application to Gene-Expression Analysis

Sepp Hochreiter
hochreit@cs.tu-berlin.de

Klaus Obermayer
Fakultät für Elektrotechnik und Informatik
Technische Universität Berlin
Franklinstr. 28/29, 10587 Berlin, Germany
oby@cs.tu-berlin.de

We consider the classification task for datasets which are described by matrices. Rows and columns of these matrices correspond to objects where row and column objects may be from different sets and column objects are labeled. Data matrix entries express relationships between row and column objects and are produced by an unknown kernel. These kernels represent dot products in some (unknown) feature space. In this feature space a linear column object classifier should be constructed. However the dot products between column objects are not available. Therefore standard support vector techniques cannot be utilized. We derive a new objective function for model selection in such a feature space according to the principle of structural risk minimization. The new objective allows the analysis of matrices which are not positive definite, and not even symmetric or square. Because row objects can be interpreted as features and our method assigns support vector weights to the row objects can be used for feature selection. An additional constraint, which imposes sparseness on the row objects resulting in few selected features. We analyse data obtained from DNA microarrays, where "column" objects correspond to samples, "row" objects correspond to genes and matrix elements correspond to expression levels. Benchmarks are conducted using standard one-gene classification and support vector machines and K-nearest neighbors after standard feature selection. Our new method extracts a sparse set of genes and provides superior classification results.

Introduction

Many data in the real world are characterized by matrices, e.g. data in Microsoft Excel tables, transaction data or micro arrays. Rows and columns of these matrices indicate the relationship between objects. One typical case are so-called pairwise data, where rows as well as columns of the matrix represent the objects of the dataset (Fig. 1a) and where the entries of the matrix denote similarity values. Another typical case occurs, if objects are described by a set of features (Fig. 1b). In this case, the column objects are the objects to be characterized, the row objects correspond to their features and the matrix elements denote the strength with which a feature is expressed in a particular object.

Pairwise Data (a)

	A	B	C	D	E	F	G	H	I	J	K	L
A	0.9	-0.1	-0.8	0.5	0.2	-0.5	-0.7	-0.9	0.2	-0.7	0.4	-0.3
B	-0.1	0.9	0.6	0.3	-0.7	-0.6	0.3	0.7	-0.3	-0.8	-0.7	-0.9
C	-0.8	0.6	0.9	0.2	-0.6	0.6	0.5	0.2	-0.7	-0.5	-0.1	0.6
D	0.5	0.3	0.2	0.9	0.7	0.1	0.3	-0.1	0.6	0.9	-0.9	-0.1
E	0.2	-0.7	-0.6	0.7	0.9	-0.9	-0.5	0.4	0.1	0.3	-0.6	0.7
F	-0.5	-0.6	0.6	0.1	-0.9	0.9	0.9	-0.2	-0.6	-0.5	-0.4	-0.3
G	-0.7	0.3	0.5	0.3	-0.5	0.9	0.9	-0.3	-0.3	0.6	0.9	-0.7
H	-0.9	0.7	0.2	-0.1	0.4	-0.2	-0.3	0.9	0.2	-0.9	0.3	0.4
I	0.2	-0.3	-0.7	0.6	0.1	-0.6	-0.3	0.2	0.9	-0.3	-0.7	0.8
J	-0.7	-0.8	-0.5	0.9	-0.3	-0.5	0.6	-0.9	-0.3	0.9	-0.1	-0.5
K	0.4	-0.7	-0.1	-0.9	-0.6	-0.4	0.9	0.3	-0.7	-0.1	0.9	0.1
L	-0.3	-0.9	0.6	-0.1	0.7	-0.3	-0.7	0.4	0.8	-0.5	0.1	0.9

Feature Vectors (b)

	A	B	C	D	E	F	G
α	1.3	-2.2	-1.6	7.8	6.6	-7.5	-4.8
β	-1.8	-1.1	7.2	2.3	9.0	3.8	3.9
χ	1.2	1.9	-2.9	-2.2	-4.4	-4.7	-8.4
δ	3.7	0.8	-0.6	2.5	-5.7	0.1	-0.3
ε	9.2	-9.4	-8.3	9.2	-2.4	-3.9	1.9
φ	-7.7	8.6	-9.7	-7.4	2.6	6.9	2.9
γ	-4.8	0.1	-1.2	0.9	0.2	2.7	0.2
η	0.7	-1.7	0.3	-7.2	-1.8	4.6	2.6
ι	-6.2	-6.2	1.8	3.6	-0.7	-9.4	0.9
φ	9.0	4.8	-8.3	-0.8	-2.0	4.4	-1.9
κ	6.2	9.0	1.5	-1.1	7.7	8.4	-2.1
λ	9.6	7.0	2.5	-4.3	5.4	0.7	1.2

Figure* 1: *Two matrix data examples: pairwise data and feature vectors (see text).

In the following we consider the task of learning a classification problem on matrix data. We consider the case that class labels are assigned to the column objects of the training set. Given the matrix and the class labels we then want to construct a classifier with good generalization properties. From all the possible choices we select classifiers from the support vector

machine (SVM) family [8, 4] and we use the principle of structural risk minimization [8] for model selection - because of its recent success [4] and its theoretical properties [8].

A serious problem arises when the number of features becomes large and comparable to the number of objects: Without feature selection, SVMs are prone to overfitting, despite the complexity regularization which is implicit in the learning method [1]. Rather than being sparse in the number of support vectors, the classifier should be sparse in the number of features used for classification. This relates to the result [8] that the number of features provide an upper bound on the number of "essential" support vectors. We show how to limit the number of features and, therefore, how to increase the generalization capability, i.e. the performance, of support vector approaches.

The New Objective And The Resulting Optimization Formulation

Given are two sets $\mathcal{X}$ and $\mathcal{Z}$ of objects, which are described by feature vectors $\boldsymbol{x}$ and $\boldsymbol{z}$. Based on the feature vectors $\boldsymbol{x}$ we construct a linear classifier defined as $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$, where $\langle ., . \rangle$ denotes a dot product. The zero isoline of f is a hyperplane with unit normal vector $\hat{\boldsymbol{w}}$ and perpendicular distance $b/\|\boldsymbol{w}\|_2$ from the origin. The hyperplane's margin γ is given by $\gamma = \min_{\boldsymbol{x} \in \mathcal{X}} |\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle + b/\|\boldsymbol{w}\|_2|$. Setting $\gamma = \|\boldsymbol{w}\|_2^{-1}$ allows us to treat normal vectors $\boldsymbol{w}$ which are not normalized, if the margin is normalized to 1. The hyperplane with largest margin is then obtained by minimizing $\|\boldsymbol{w}\|_2^2$ for a margin which equals 1. It has been shown [7, 6, 5] that the generalization error of a linear classifier can be bounded from above with probability $1 - \delta$ by the bound $\mathcal{B}$,

$$(1) \qquad \mathcal{B}(L, a/\gamma, \delta) = \frac{2}{L} \left(\log_2 \left(E\mathcal{N} \left(\frac{\gamma}{2\ a}, \mathcal{F}, 2L \right) \right) + \log_2 \left(\frac{4\ L\ a}{\delta\ \gamma} \right) \right) ,$$

provided that the training classification error is zero and $f(\boldsymbol{x})$ is bounded by $-a \leq f(\boldsymbol{x}) \leq a$ for all $\boldsymbol{x}$ drawn iid. L denotes the number of training objects $\boldsymbol{x}$ and $E\mathcal{N}(\epsilon, \mathcal{F}, L)$ the expected ϵ-covering number of a class $\mathcal{F}$ of functions that map data objects from T to $[0, 1]$ (see [5]). To select a classifier with good generalization properties we suggest to minimize a/γ under proper constraints. a is not known in general. Therefore we approximate a by the range $m = 0.5\ (\max_i \langle \hat{\boldsymbol{w}}, \boldsymbol{x}^i \rangle - min_i \langle \hat{\boldsymbol{w}}, \boldsymbol{x}^i \rangle)$ of values in the training set and minimize $\mathcal{B}(L, m/\gamma, \delta)$ instead of eq. (1).

Let $\boldsymbol{X} := \left(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^L \right)$ be the matrix of feature vectors of L objects from the set $\mathcal{X}$ and $\boldsymbol{Z} := \left(\boldsymbol{z}^1, \boldsymbol{z}^2, \ldots, \boldsymbol{z}^P \right)$ be the matrix of feature vectors of P objects from the set $\mathcal{Z}$. The objects of set $\mathcal{X}$ are labeled, and we summarize all labels using a diagonal label matrix $\boldsymbol{Y}$. The feature vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$ are unknown, but the matrix $\boldsymbol{K} := \boldsymbol{X}^T\ \boldsymbol{Z}$ of the corresponding scalar products is given. The training set is $\boldsymbol{K}, \boldsymbol{Y}$. The principle of structural risk minimization is implemented by minimizing an upper bound on $(m/\gamma)^2$ given by $\|\boldsymbol{X}^T\ \boldsymbol{w}\|_2^2$, as can be seen from $m/\gamma \leq \|\boldsymbol{w}\|_2\ \max_i |\langle \hat{\boldsymbol{w}}, \boldsymbol{x}^i \rangle| \leq \sqrt{\sum_i (\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle)^2} = \|\boldsymbol{X}^T\ \boldsymbol{w}\|_2$. In the following optimization formulation the constraints $f(\boldsymbol{x}^i) = y^i$ imposed by the training set are taken into account including slack variables $\xi_i^+, \xi_i^- \geq 0$:

$$(2) \qquad \begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-} \quad & \frac{1}{2}\ \|\boldsymbol{X}^T\ \boldsymbol{w}\|_2^2 + M^+\ \mathbf{1}^T \boldsymbol{\xi}^+ + M^-\ \mathbf{1}^T \boldsymbol{\xi}^- \\ \text{s.t.} \quad & \boldsymbol{Y}^{-1} \left(\boldsymbol{X}^T\ \boldsymbol{w} + b\mathbf{1} \right) - \mathbf{1} + \boldsymbol{\xi}^+ \geq \mathbf{0} \\ & \boldsymbol{Y}^{-1} \left(\boldsymbol{X}^T\ \boldsymbol{w} + b\mathbf{1} \right) - \mathbf{1} - \boldsymbol{\xi}^- \leq \mathbf{0}\ ,\ \boldsymbol{\xi}^+, \boldsymbol{\xi}^- \geq \mathbf{0}\ . \end{aligned}$$

M^+ penalizes wrong classification and for classification M^- may be set to zero.

Let $\tilde{\boldsymbol{\alpha}}^+, \tilde{\boldsymbol{\alpha}}^-$ be the dual variables for the constraints imposed by the training set, $\tilde{\boldsymbol{\alpha}} := \tilde{\boldsymbol{\alpha}}^+ - \tilde{\boldsymbol{\alpha}}^-$, and $\boldsymbol{\alpha}$ a vector with $\tilde{\boldsymbol{\alpha}} = \boldsymbol{Y} \left(\boldsymbol{X}^T\ \boldsymbol{Z} \right) \boldsymbol{\alpha}$. Two cases must be treated: $\boldsymbol{\alpha}$ is not unique

and $\boldsymbol{\alpha}$ does not exist. If $\boldsymbol{\alpha}$ is not unique we choose $\boldsymbol{\alpha}$ which is most sparse in its components (see below). If $\boldsymbol{\alpha}$ does not exist we set $\boldsymbol{\alpha} = \left(\boldsymbol{Z}^T \boldsymbol{X} \boldsymbol{Y}^{-T} \boldsymbol{Y}^{-1} \boldsymbol{X}^T \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T \boldsymbol{X} \boldsymbol{Y}^{-T} \tilde{\boldsymbol{\alpha}}$, where $\boldsymbol{Y}^{-T} \boldsymbol{Y}^{-1}$ is the identity. We derive for the optimal values: $\boldsymbol{w} = \boldsymbol{Z} \boldsymbol{\alpha}$, $0 = \mathbf{1}^T \left(\boldsymbol{X}^T \boldsymbol{Z}\right) \boldsymbol{\alpha}$, $\tilde{\alpha}_i \leq M^+$, and $-\tilde{\alpha}_i \leq M^-$. For $M^+ = M^- = M$ and $C := M \, \|\boldsymbol{Y} \left(\boldsymbol{X}^T \boldsymbol{Z}\right)\|_{row}^{-1}$ we get $\|\boldsymbol{\alpha}\|_\infty \leq C$ and $\|\tilde{\boldsymbol{\alpha}}\|_\infty \leq M$, where $\|.\|_{row}$ is the row-sum norm. With an additional sparseness term $\epsilon \|\boldsymbol{\alpha}\|_1$, we obtain the dual problem:

$$(3) \qquad \begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right)^T \boldsymbol{K}^T \boldsymbol{K} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) - \mathbf{1}^T \boldsymbol{Y} \boldsymbol{K} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) + \epsilon \, \mathbf{1}^T \left(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-\right) \\ \text{s.t.} & \mathbf{1}^T \boldsymbol{K} \left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) = 0 \, , \; C\mathbf{1} \geq \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^- \geq \mathbf{0} \; . \end{array}$$

If a classifier has been selected according to eq. (3), a new example $\boldsymbol{u}$ is classified according to the sign of $f(\boldsymbol{u}) = \langle \boldsymbol{w}, \boldsymbol{u} \rangle + b = \sum_{i=1}^{P} \alpha_i \langle \boldsymbol{z}^i, \boldsymbol{u} \rangle + b$.

The optimal classifier is selected by optimizing eq. (3), and as long as $a = m$ holds true for all possible objects $\boldsymbol{x}$ (which are assumed to be drawn iid), the generalization error is bounded by eq. (1). If outliers are rejected, condition $a = m$ can always be enforced. For large training sets the number of rejections is small: The probability $P\{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle| > m\}$ that an outlier occurs can be bounded with confidence $1 - \delta$ using the additive Chernoff bounds (e.g. [8]): $P\{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle| > m\} \leq \sqrt{\frac{-\log \delta}{2L}}$. But note, that not all outliers are misclassified, and the trivial bound on the generalization error is still of the order L^{-1}.

Kernel Functions Are Dot Products

In the last section we have assumed that the matrix $\boldsymbol{K}$ is derived from scalar products between the feature vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ which describe the objects from the sets $\mathcal{X}$ and $\mathcal{Z}$. For all practical purposes, however, the only information available is summarized in the matrices $\boldsymbol{K}$ and $\boldsymbol{Y}$. The feature vectors are not known and it is even unclear whether they exist. In order to apply the results of previous section to practical problems the following question remains to be answered: What are the conditions under which the measurement operator $k(., \boldsymbol{z})$ can indeed be interpreted as a scalar product between feature vectors and under which the matrix $\boldsymbol{K}$ can be interpreted as a Gram matrix?

The following theorem states under which assumptions on $k(., \boldsymbol{z})$ this kernel represents a dot product. Let $L^2(H)$ denote the set of functions h from H with $\int h^2(\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Theorem 1 (Singular Value Expansion) *Let H_1 and H_2 be Hilbert spaces. Let α be from $L^2(H_1)$ and let k be a kernel from $L^2(H_2, H_1)$ which defines a Hilbert-Schmidt operator $T_k : H_1 \to H_2$ $(T_k \alpha)(\boldsymbol{x}) = f(\boldsymbol{x}) = \int k(\boldsymbol{x}, \boldsymbol{z}) \, \alpha(\boldsymbol{z}) \, d\boldsymbol{z}$.*
Then there exists an expansion $k(\boldsymbol{x}, \boldsymbol{z}) = \sum_n s_n \, e_n(\boldsymbol{z}) \, g_n(\boldsymbol{x})$ which converges in the L^2-sense. The $s_n \geq 0$ are the singular values of T_k, and $e_n \in H_1$, $g_n \in H_2$ are the corresponding orthonormal functions.

DNA Microarray Data Analysis

We apply our new method eq. (3) to the DNA microarray data published in [3]. Column objects are samples from different brain tumors of the medullablastoma kind. The samples were obtained from 60 patients, which were treated in a similar way and the samples were labeled according to whether a patient responded well to chemo- or radiation therapy. Row objects correspond to genes. For every sample-gene pair a snapshot of the level of gene expression was measured. This gave rise to a $60 \times 7,129$ real valued sample-gene matrix.

The task is now to construct a classifier which predicts therapy outcome on the basis of samples taken from new patients. The major problem of this classification task is the

limited number of samples - given the large number of genes. Therefore, feature selection is a prerequisite for good generalization [2]. For feature selection we apply our new method on a $59 \times 7,129$ matrix, where one column object was withhold to avoid biased feature selection.

Table 1: ***Prediction of the outcome of chemo therapy based on gene-expression. Our method (P-SVM) is compared to standard methods via the number of wrong classifications ("E"). Standard method results are taken from [3].***

Meth	F	E	Meth	C	F	E	Meth	C	F	E
TrkC	1	**20**	KNN		8	**13**	P-SVM	0.1	25/30/35	**8/5/8**
SVM		**15**	C2			**12**	P-SVM	0.15	25/30/35	**8/5/7**
C1		**14**	P-SVM	0.05	25/30/35	**8/5/8**	P-SVM	0.2	25/30/35	**6/4/6**

Table 1 shows benchmark results for DNA microarray data where the classification error given by the number of wrong classifications ("E") for different numbers of selected features ("F") and for different values of the parameter C. Data are provided for "TrkC"-Gene classification, standard SVMs, weighted "TrkC"/SVM (C1), K nearest neighbor (KNN), combined SVM/TrkC/KNN (C2), and our procedure (P-SVM). Standard methods feature selection was based on the correlation of features with classes using signal-to-noise-statistics and t-statistics according to [1]. The feature selection procedure (also a classifier) had its lowest misclassification rate between 20 and 40 features. Our method clearly outperforms standard methods — the number of misclassification is down by a factor of 2 (for 30 selected genes) using p-SVM.

REFERENCES

[1] R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46:389–422, 2002.

[3] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[4] B. Schölkopf and A. J. Smola. *Learning with kernels — Support Vector Machines, Reglarization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.

[5] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anhtony. A framework for structural risk minimisation. In *Comp. Learn. Th.*, pages 68–76, 1996.

[6] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anhtony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44:1926–1940, 1998.

[7] J. Shawe-Taylor and N. Cristianini. On the generalisation of soft margin algorithms. Technical Report NC2-TR-2000-082, Dep. Computer Science, Royal Holloway, London, 2000.

[8] V. Vapnik. *The nature of statistical learning theory*. Springer, NY, 1995.