CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY

REPRINT

Kernel represenation of the Kesler construction for Multi-class SVM classification

Vojtěch Franc, Václav Hlaváč

xfrancv@cmp.felk.cvut.cz

Vojěch Franc and Václav Hlaváč. Kernel Representation of the Kesler Construction for Multi-class SVM Classification. In H. Wildenauer and W. Kropatsch, editors, Proceedings of the CVWW'02, page 7, Wien, Austria, February 2002. PRIP.

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/franc/franc-multiKernel02.pdf

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Kernel represenation of the Kesler construction for Multi-class SVM classification

Vojtěch Franc, Václav Hlaváč

Abstract

We propose a transformation from the multi-class SVM classification problem to the single-class SVM problem which is more convenient for optimization. The proposed transformation is based on simplifying the original problem and employing the Kesler construction which can be carried out by the use of properly defined kernel only. The experiments conducted indicate that the proposed method is comparable with the one-against-all decomposition solved by the state-of-the-art SMO algorithm.

1 Introduction

The standard Support Vector Machines (SVM) [8] are designed for dichotomic classification problem (two classes only, called also binary classification). The multi-class classification problem is commonly solved by a decomposition to several binary problems for which the standard SVM can be used. For instance, one-against-all (1-a-a) decomposition is often applied. In this case the classification problem to k classes is decomposed to k dichotomic decisions $f_m(x)$, $m \in K = \{1, \ldots, k\}$, where the rule $f_m(x)$ separates training data of the m-th class from the other training patterns. The classification of a pattern x is performed according to maximal value of functions $f_m(x)$, $m \in K$, i.e., the label of x is computed as $\mathrm{argmax}_{m \in K} f_m(x)$.

For the SVM, however, the multi-class problem can be solved directly [8, 9]. Let us consider that we are given labelled training patterns $\{(x_i, y_i) \colon i \in I\}$, where a pattern x_i is from an n-dimensional space $\mathcal{X}$ and its label attains a value from a set K. The $I = \{1, \ldots, l\}$ denotes a set of indices. The linear classification rules $f_m(x) = \langle w_m, x \rangle + b_m$, $m \in K$ (the dot product is denoted by $\langle ., . \rangle$) can be found directly by solving the multi-class SVM problem

$$\begin{aligned} &\min_{w,b,\xi} \tfrac{1}{2} \sum_{m \in K} ||w_m||^2 + C \cdot \sum_{i \in I} \sum_{m \in K \setminus \{y_i\}} (\xi_i^m)^d \,, \\ \text{s.t.} \quad &\langle w_{y_i}, x_i \rangle + b_{y_i} - (\langle w_m, x_i \rangle + b_m) \geq 1 - \xi_i^m \,, \\ &\xi_i^m \geq 0 \,, \quad i \in I \,, m \in K \setminus \{y_i\} \,. \end{aligned} \tag{1}$$

Similarly to the dichotomic SVM, the minimization of the sum of norms $\|w_m\|^2$ leads to maximization of the margin between classes. For a non-separable case, the sum of $(\xi_i^m)^d$ weighted by a regularization constant C means that the cost function penalizes misclassification of training data. The linear $(d = 1)$ or quadratic $(d = 2)$ cost functions are often used.

To employ kernel functions [8] into non-linear classification rules $f_m(x)$, one has to formulate a dual form of the multi-class SVM decision (1) which is defined as [8, 9]

$$\min_{\alpha} \sum_{i\in I}\sum_{j\in I}(\tfrac{1}{2}c_j^{y_i}A_iA_j - \sum_{m\in K}\alpha_i^m\alpha_j^{y_i} + \tfrac{1}{2}\sum_{m\in K}\alpha_i^n\alpha_j^m)k(x_i,x_j) - 2\sum_{i\in I}\sum_{m\in K}\alpha_i^m \,,$$

$$\begin{gathered} \text{s.t.} \quad \sum_{i\in I}\alpha_i^m = \sum_{i\in I}c_i^m A_i \,, m\in K \,, \\ 0 \le \alpha_i^m \le C \,, \alpha_i^{y_i} = 0, \\ A_i = \sum_{m\in K}\alpha_i^m \,, c_j^{y_i} = \begin{cases} 1 & \text{if} \quad y_i = y_j \,, \\ 0 & \text{if} \quad y_i \ne y_j \,, \end{cases} \\ i\in I \,, m\in K \,. \end{gathered} \tag{2}$$

The dual problem (2) has $k\cdot l$ variables and l of them are always zero. The number of variables is too large in practical problems and consequently it is very difficult to solve the dual quadratic problem directly. There is a solution which employs a decomposition method and solves series of smaller quadratic problems. However, the constraints of the problem (2) are too complicated to allow direct use of efficient decomposition methods developed for dichotomic decision problems, e.g., the Sequential Minimal Optimizer (SMO) algorithm [6].

We propose (i) to modify slightly the original problem (1) by adding the term $(1/2)\sum_{m\in K} b_m^2$ to the objective function, and (ii) to transform the modified problem to the single-class SVM problem which is considerably simpler than the previous formulation. Efficient algorithms can be used to solve the new problem. Moreover, the proposed transformation can be performed by the properly defined kernel function only. The addition of the $(1/2)\,b$ term in the objective function was suggested by Mangasarian [5] for the dichotomic problem. Solutions of the modified problem mostly coincides with the solutions of the original problem [5].

The following section describes proposed approach in details.

2 From multi-class SVM to single-class SVM

We consider modified multi-class SVM where the $(1/2)\,b^2$ is added to the objective function of the (1) which leads to

$$\min_{w,b,\xi} \tfrac{1}{2}\sum_{m\in K}(||w_m||^2 + b^2) + C\cdot\sum_{i\in I}\sum_{m\in K\setminus\{y_i\}}(\xi_i^m)^d \,,$$

$$\begin{gathered} \text{s.t.} \\ \langle w_{y_i}, x_i\rangle + b_{y_i} - (\langle w_m, x_i\rangle + b_m) \ge 1 - \xi_i^m \,, \\ \xi_i^m \ge 0 \,, \quad i\in I \, m\in K\setminus\{y_i\} \,. \end{gathered} \tag{3}$$

We name the problem (3) defined above as the multi-class BSVM problem (B stands for the added bias). Next we introduce a transformation which translates the multi-class BSVM prob-

lem (3) to the single-class SVM problem. The single-class SVM problem is defined as

$$\begin{aligned} &\min_{w,\xi} \tfrac{1}{2}||w||^2 + C\sum_{i\in I}(\xi_i)^d \,, \\ &\text{s.t.} \quad \langle w, z_i\rangle \geq 1-\xi_i \,, \quad i \in I \,. \end{aligned} \tag{4}$$

This problem (4) can be already solved by algorithms which are considerably simpler than the original problems (1) or (3). The dual form of the problem (4) with the linear cost function $d = 1$ is

$$\begin{aligned} &\max_{\alpha} \sum_{i\in I}\alpha_i - \tfrac{1}{2}\sum_{i\in I}\sum_{j\in I}\alpha_i\cdot\alpha_j\cdot k(z_i,z_j) \,, \\ &\text{s.t.} \qquad 0 \leq \alpha_i \leq C \,, \quad i \in I \,, \end{aligned} \tag{5}$$

where $k(z_i, z_j)$ was substituted for the dot products $\langle z_i, z_j\rangle$. The case with the quadratic cost function $d = 2$ can be solved as the separable case using the kernel function $k'(x_i, x_j) = k(x_i, x_j) + \delta_{i,j}\cdot\frac{1}{2C}$. The dual form of the separable case is the same as the problem (5) up to the constraints which simplify to $0 \leq \alpha_i$. We will describe two simple algorithms for solving the single-class SVM problem in Section 3.

The transformation from the multi-class BSVM problem to the single-class SVM problem is based on the Kesler's construction [1]. This construction maps the input n-dimensional space $\mathcal{X}$ to a new $(n+1)\cdot k$-dimensional space $\mathcal{Y}$ where the multi-class problem appears as the single-class problem. Each training pattern x_i is mapped to new $(k-1)$ patterns z_i^m, $m \in K \setminus \{y_i\}$ defined as follows. Let us assume that coordinates of z_i^m are divided into k slots. If each slot $z_i^m(j)$, $j \in K$ has $n+1$ coordinates then

$$z_i^m(j) = \begin{cases} [x_i, 1] \,, & \text{for} \quad j = y_i \,, \\ -[x_i, 1] \,, & \text{for} \quad j = m \,, \\ 0 \,, & \text{otherwise.} \end{cases} \tag{6}$$

We seek a vector w composed of vectors $w_1, \ldots, w_k$ and thresholds $b_1, \ldots, b_k$ in the new space $\mathcal{Y}$ as

$$w = [[w_1, b_1], [w_2, b_2], \ldots, [w_k, b_k]] \,. \tag{7}$$

For instance, when $k = 4$ and $y_i = 3$ then the vectors z_i^m, $m = 1, 2, 4$ are constructed as

$$\begin{array}{rcccccc} z_i^1 & = & [& -[x_i,1] & 0 & [x_i,1] & 0 &] \\ z_i^2 & = & [& 0 & -[x_i,1] & [x_i,1] & 0 &] \\ z_i^4 & = & [& 0 & 0 & [x_i,1] & -[x_i,1] &] \end{array}$$

Performing the transformation (6) we obtain a set $\{z_i^m : i \in I\, m \in K\setminus\{y_i\}\}$ containing $(k-1)\cdot l$ vectors. Each constraint of the multi-class BSVM problem can be expressed as $\langle w, z_i^m\rangle \geq 1 - \xi_i^m$ using the transformed vectors. It is obvious that by substituting w to the objective function of the single-class SVM problem the objective function (4) becomes equivalent to the

objective function (3) of the multi-class BSVM. Consequently, the multi-class BSVM problem can be equivalently expressed as the single-class SVM problem,

$$\begin{array}{l} \min\limits_{w} \frac{1}{2}||w||^2 + C \cdot \sum\limits_{i \in I} \sum\limits_{m \in K \setminus \{y_i\}} (\xi_i^m)^d \,, \\ \text{s.t.} \quad \langle w, z_i^m \rangle \geq 1 - \xi_i^m \,, \\ \quad i \in I \, m \in K \setminus \{y_i\} \,. \end{array} \tag{8}$$

At a first look the introduced transformation seems to be intractable because of increased dimension. However, in the dual form in which the data appears in terms of dot products only the transformation can be performed by introducing a properly defined kernel function.

Let z_i^m and z_j^n be two vectors from $\mathcal{Y}$ created by the transformation (6). Note that the vector z_i^m has the y_i-th coordinate slot equal to $[x_i, 1]$, the m-th slot equal to $-[x_i, 1]$, and remaining coordinates equal to zero. The vector z_j^n is created likewise. Consequently, the dot product $\langle z_i^m, z_j^n \rangle$ is equal to the sum of dot products between $[x_i, 1]$ and $[x_j, 1]$ which occupy the same coordinate slot. The sign of these dot products is positive if $y_i = y_j$ or $m = n$ and negative if $y_i = n$ or $y_j = n$. If all the numbers y_i, y_j, m, and n differ then the dot product is equal to zero. The construction of the dot product $\langle z_i^m, z_j^n \rangle$ can be easily expressed using the Kronecker delta, i.e., $\delta(i, j) = 1$ for $i = j$, and $\delta(i, j) = 0$ for $i \neq j$. The dot product between z_i^m and z_j^n is

$$\langle z_i^m, z_j^n \rangle = (\langle x_i, x_j \rangle + 1) \cdot (\delta(y_i, y_j) + \delta(m, n) - \delta(y_i, n) - \delta(y_j, m)) \,.$$

The dot products $\langle x_i, x_j \rangle$ are replaced by the kernel function $k(x_i, x_j)$ in the non-linear case. The kernel function $k'(z_i^m, z_j^n)$ involving transformations (6) and non-linear case is constructed as

$$k'(z_i^m, z_j^n) = (k(x_i, x_j) + 1) \cdot (\delta(y_i, y_j) + \delta(m, n) - \delta(y_i, n) - \delta(y_j, m)) \,. \tag{9}$$

It implies that solving the dual form (5) of the single-class SVM problem with the kernel (9) is equivalent to solving the dual form of the multi-class BSVM problem (3). As the result of the dual single-class problem we obtain a set of α_i^m, $i = 1, \ldots$, $m = 1, \ldots, k$, $m \neq y_i$ multipliers corresponding to the transformed vectors z_i^m. These multipliers α_i^m determine the vectors w_m and thresholds b_m which can be obtained by reverting the transform (7).

The normal vector w in the transformed space $\mathcal{Y}$ is equal to $w = \sum_{i \in I} \sum_{m \in K \setminus \{y_i\}} z_i^m \alpha_i^m$. The vector $w_j \in \mathcal{X}$ occupies the j-th coordinate slot and is determined by the weighted sum of vectors z_i^m which have the non-zero j-th coordinate slot, so that

$$\begin{aligned} w_j &= \sum_{i \in I} \sum_{m \in K \setminus \{y_i\}} x_i \alpha_i^m (\delta(j, y_i) - \delta(j, m)) \,, \\ b_j &= \sum_{i \in I} \sum_{m \in K \setminus \{y_i\}} \alpha_i^m (\delta(j, y_i) - \delta(j, m)) \,, \end{aligned}$$

holds. To classify the pattern x in the non-linear case there is need to evaluate $f_j = \langle w_j, \phi(x) \rangle + b_j$ which is equal to

$$f_j(x) = \sum_{i \in I} k(x_i, x) \sum_{m \in K \setminus \{y_i\}} \alpha_i^m (\delta(j, y_i) - \delta(j, m)) + b_j \,.$$

Table 1: Benchmark datasets used for testing.

	number of patterns	number of classes	number of attributes
iris	150	3	4
wine	178	3	13
glass	214	6	13
thyroid	215	3	3

3 Algorithms to the single-class SVM problem

The introduced kernel allows us to solve the multi-class BSVM problem by the use of algorithms solving the single-class SVM problem. Many efficient optimization algorithms for the two-class problem can be readily modified to solve the one-class problem. We have conducted several experiments (see Section 4) using the modified Sequential Minimal Optimizer (SMO) [6] and the kernel Schlesinger-Kozinec algorithm [3].

The SMO for the single-class SVM problem can modify only one Lagrangian at a time since the dual form does not contain the equality constrains. The framework of the modified algorithm is preserved from the original one.

The kernel Schlesinger-Kozinec algorithm solves the two-class SVM problem with quadratic cost function. This problem is transformed to the equivalent problem where the nearest points from the convex hulls are sought. This transformed probllem can be solved by a simple iterative procedure. The nearest point from the origin to one convex hull is sought in the modification to the single-class SVM problem. We used the modified kernel Schlesinger-Kozinec's algorithm to train the multi-class BSVM problem with quadratic cost function and the modified SMO algorithm for the linear cost function.

The implementation of both algorithms in Matlab is available [2].

4 Experiments

We tested the proposed method on the benchmark data sets selected from the UCI data repository [7] and Statlog data collection. We scaled all the data to range $[-1, 1]$. Table 1 summarizes the data sets used.

As a comparative approach we used the one-against-all decomposition and the SMO [6] algorithm for learning the decomposed dichotomic SVM problems which we denote 1-a-a SMO. To solve the single-class problem obtained employing the proposed kernel we used (i) the simplified SMO algorithm denoted as M-1-SMO and (ii) the kernel Schlesinger-Kozinec algorithm denoted as M-1-KSK both mentioned in Section 3.

We trained the classifiers using the Radial Basis Function (RBF) kernel $k(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{2 \cdot \sigma}}$ with the $\sigma = \{2^{-3}, 2^{-2}, \ldots, 2^3\}$ and the regularization constant $C = \{2^0, 2^1, \ldots, 2^7\}$. Each

Table 2: Results of comparison on the benchmark datasets. Measured: testing classification error **CE** $[\%]$, training **time** $[s]$ and number of support vectors **SVs**.

		1-a-a SMO	M-1-SMO	M-1-KSK
iris	CE (C, σ)	2.7 $(2^7, 2^0)$	**2.0**$(2^5, 2^0)$	**2.0** $(2^4, 2^0)$
	time	**0.12**	0.22	0.44
	SVs	**17**	30	19
wine	CE (C, σ)	**1.1** $(2^5, 2^3)$	2.3 $(2^6, 2^3)$	1.7 $(2^1, 2^1)$
	time	**0.2**	0.67	0.40
	SVs	54	**37**	54
glass	CE(C, σ)	37.0 $(2^5, 2^{-1})$	**28.7** $(2^3, 2^{-2})$	31.1 $(2^0, 2^{-2})$
	time	14.10	4.06	**1.37**
	SVs	**150**	167	177
thyroid	CE(C, σ)	2.3 $(2^4, 2^0)$	2.7 $(2^1, 2^{-1})$	**1.8**$(2^0, 2^{-1})$
	time	0.41	**0.13**	0.31
	SVs	**35**	43	66

from the 7×8 pairs of (σ, C) was evaluated using 10-fold cross validation method. The parameters which yielded the best average testing error rate are enlisted in Table 2. We also measured average values of (i) the number of support vectors and (ii) the training time on training time on Pentium PIII/750Mhz and (ii) number of support vectors.

5 Conclusions and future work

We propose a transformation from the multi-class SVM classification problem (1) to the single-class SVM problem (4) for which efficient optimization algorithms exist. First the original problem is slightly modified by adding the term $(1/2)\sum_{m \in K} b_m^2$ (similarly to Mangasarian [5] in the dichotomic problem). Then the modified problem is transformed to the single-class SVM problem which is carried out by the use of a properly defined kernel function only.

The experiments conducted indicate that the proposed method is comparable with the one-against-all decomposition solved by the state-of-the-art SMO algorithm. It is worthwhile to investigate the proposed kernel with other efficient algorithms which can solve the single-class problem, e.g. the Nearest Point Algorithm [4] or the Successive Overrelaxation (SOR) algorithm [5].

Acknowledgements

Our research was by the European Union under project IST-2001- 32184, by the Czech Ministry of Education under projects MSM 212300013, MSMT Kontakt ME412, and by the Grant Agency of the Czech Republic under project GACR 102/00/1679.

References

[1] O. Duda, R., E. Hart, P., and G. Stork, D. *Pattern Recognition*. John Wiley & Sons, 2000.

[2] V. Franc and V. Hlavač. Statistical pattern recognition toolbox for Matlab, 2000-2001. `http://cmp.felk.cvut.cz`.

[3] Vojtěch Franc and Václav Hlaváč. A simple learning algorithm for maximal margin classifier. In A. Leonardis and H. Bischof, editors, *Kernel and Subspace Methods for Computer Vision*, pages 1–11, Vienna, Austria, August 2001. TU Vienna.

[4] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Transactions on Neural Networks*, 11(1):124–136, January 2000.

[5] L. Mangasarian, O. and R. Musicant, D. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10(5), 1999.

[6] J.C. Platt. Fast training of support vectors machines using sequential minimal optimization. In B. Scholkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods*. MIT Press, Cambridge, MA., USA, 1998.

[7] UCI-benchmark repository of artifical and real data sets. University of California Irvine, `http://www.ics.uci.edu/~mlearn`.

[8] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[9] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX, UK, 1998.