ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 161-166

An Unified Framework for 'All data at once' Multi-Class Support Vector Machines

Cecilio Angulo, Xavier Parra, Andreu Català

Technical University of Catalonia, GREC Research Group,
Av. Víctor Balaguer s/n, Vilanova i la Geltrú, Spain

Abstract

Support Vectors (SV) are a machine learning procedure based on Vapnik's Statistical Learning Theory, initially defined for bi-classification problems. A lot of work is being made from different research areas to obtain new algorithms for multi-class problems, the more usual task in real-world problems. A promising extension is to treat 'all data at once' into one multi-class SVM by modifying the associated quadratic programming (QP) problem. In this work, a unified architecture is developed to compare the associated QP problem for different approaches. With the new framework comparisons between algorithms become easier and it is a powerful tool to analyze the performance and behaviour of these approaches.

1 Introduction

Support Vector Machines are learning machines implementing the structural risk minimization inductive principle to obtain good generalization on a limited number of learning patterns. This theory was originally developed by Vapnik on the basis of a separable bipartition problem [7]. New algorithms have been derived in the past years for categorization problems with no separable data and regression problems. For the multi-class classification problem, there exists, roughly speaking, two algorithmic approaches. The first approach proposes to divide the initial multi-class problem treating K classes into K sub-problems, each having in consideration a bipartition of one class versus the rest of classes [7]. Another possibility is to create $L = K(K-1)/2$ bipartition dividing each class from another one [5]. As extensions, it is possible to use error correcting output codes (ECOC) to select the classes for the bipartition [3], or a ternary partition based on a mixed formulation of SVMs for classification and regression, obtaining more robust classification [1].

The second approach is referred to develop an algorithm considering 'all data at once' [2, 4, 8]. Some structures have been constructed independently by using different initial results. This paper will show that all of these architectures can

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 161-166

be summarized into a general associated QP problem. This formulation allows to make easy comparisons between the different approaches, offering a new tool to study the performance of these algorithms and it is an open structure to propose other possibilities.

Section 2 will be devoted to formulate the structures to be considered. In Section 3 the new formulation is naturally introduced and a first result is presented making a comparison between these architectures. Finally, future work to be completed and some conclusions are presented.

2 'All data at once' SVMs

Let $\mathcal{T} = \{(\mathbf{x}_p, y_p)\}_{p=1}^{\ell} \subset \mathcal{X} \times \mathcal{Y} = \mathcal{X} \times \{\theta_1, \ldots, \theta_{K>2}\}$, be the training data set for a multi-class classification problem. A mapping $f(\mathbf{x}, \omega) = \langle \omega, \phi(\mathbf{x}) \rangle_{\mathcal{F}} + b = k(\omega, \mathbf{x}) + b$, with the smallest discrepancy with the real system answer must be chosen, where $\langle \cdot, \phi(\mathbf{x}) \rangle_{\mathcal{F}}$, denotes the inner product of $\phi(\mathbf{x})$ and another element in the features space $\mathcal{F}$, being $\phi : \mathcal{X} \to \mathcal{F}$, a non linear mapping from the original input space to a usually high dimensional space. $b = 0$ will be considered, a minimal restriction into the features space. From now on, learning machines solving this general problem considering 'all data at once' will be called KSVMCs (SVM for K-classes Classification).

2.1 A Natural Extension of SVM for Classification

The learning machine presented in this section has been independently proposed by [8] and [7] among others. Following notation in [8], if output classes are denoted only by its subscript, $\mathcal{Y} = \{\theta_1, \ldots, \theta_{K>2}\} \Rrightarrow \mathcal{Y} = \{1, \ldots, K\}$, the usual QP problem associated to a SVM for Classification can be generalized as

$$\arg\min R_{Ext}(\omega, \xi) = \frac{1}{2} \sum_{m=1}^{K} \|\omega_m\|_{\mathcal{F}}^2 + C \sum_{i=1}^{\ell} \sum_{m \neq y_i} \xi_i^{(m)}, \tag{1}$$

subject to[1]

$$\langle \omega_{y_i} - \omega_m, \phi(\mathbf{x}_i) \rangle_{\mathcal{F}} \geq 2 - \xi_i^{(m)}, \tag{2}$$

$$\xi_i^{(m)} \geq 0, \quad i = 1, \ldots, \ell \quad m \in \mathcal{Y} \backslash y_i. \tag{3}$$

Introducing Lagrange multipliers $\alpha_i^{(m)}, \beta_i^{(m)} \geq 0$, if dummy variables[2]

$$\xi_i^{(y_i)}, \alpha_i^{(y_i)}, \beta_i^{(y_i)} = 0, \quad i = 1, \ldots, \ell \tag{4}$$

are added[3], the Wolfe's dual optimization problem is obtained as follows

[1] In [7], constraint (2) is replaced by $\langle \omega_{y_i} - \omega_m, \phi(\mathbf{x}_i) \rangle_{\mathcal{F}} \geq 1 - \xi_i^{(m)}$.

[2] In [8] is considered $\xi_i^{(y_i)} = 2$, despite this assumption add a term $2C\ell$ in (1).

[3] In fact, it is only necessary to keep $\beta_i^{(y_i)}$ bounded in (5), because $\xi_i^{(y_i)} = 0$.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 161-166

$$\arg\min L_{Ext}\left(\omega,\xi,\alpha,\beta\right)=\frac{1}{2}\sum_{m=1}^{K}\left\|\omega_m\right\|_{\mathcal{F}}^2+C\sum_{i=1}^{\ell}\sum_{m=1}^{K}\xi_i^{(m)}-$$
$$-\sum_{i=1}^{\ell}\sum_{m=1}^{K}\alpha_i^{(m)}\left[\left\langle\omega_{y_i}-\omega_m,\phi\left(\mathbf{x}_i\right)\right\rangle_{\mathcal{F}}-2+\xi_i^{(m)}\right]-\sum_{i=1}^{\ell}\sum_{m=1}^{K}\beta_i^{(m)}\xi_i^{(m)}\,, \tag{5}$$
$$\xi_i^{(m)},\alpha_i^{(m)},\beta_i^{(m)}\geq 0\,,\quad i=1,\ldots,\ell\quad m\in\mathcal{Y}\backslash y_i\,.$$

By defining $c_i^{(n)}=\begin{cases}1 & \text{if } y_i=n\\ 0 & \text{if } y_i\neq n\end{cases}$, $A_i=\sum_{m=1}^{K}\alpha_i^{(m)}$, dual formulation is established as

$$W_{Ext}\left(\alpha\right)=2\sum_{i=1}^{\ell}A_i+$$
$$+\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\sum_{m=1}^{K}\left(-\frac{1}{2}c_j^{(y_i)}A_iA_j+\alpha_i^{(m)}\alpha_j^{(y_i)}-\frac{1}{2}\alpha_i^{(m)}\alpha_j^{(m)}\right)\left\langle\phi\left(\mathbf{x}_i\right),\phi\left(\mathbf{x}_j\right)\right\rangle_{\mathcal{F}}\,, \tag{6}$$

subject to $0\leq\alpha_i^{(m)}\leq C\,,\quad \alpha_i^{(y_i)}=0\,,\quad i=1,\ldots,\ell\quad m\in\mathcal{Y}\backslash y_i\,.$

The complete decision function can be written as

$$f\left(\mathbf{x}\right)=\underset{m=1,\ldots,K}{\arg\max}\left[\sum_{i:y_i=m}A_ik\left(\mathbf{x}_i,\mathbf{x}\right)-\sum_{i:y_i\neq m}\alpha_i^{(m)}k\left(\mathbf{x}_i,\mathbf{x}\right)\right]\,. \tag{7}$$

2.2 A KSVMC Based on a Uniform Convergence Result

[4] introduces a novel KSVMC structure based on the large numbers strong uniform law as an extension of the structural functional risk generalization bound for multi-class optimal hyperplanes.

Modifying the original notation in [4] with the former subscripts' notation, the generalization ability of the optimal hyperplane is improved according to the authors if the modified optimization problem

$$\arg\min R_{Conv}\left(\omega,\xi\right)=\frac{1}{2}\sum_{m=1}^{K-1}\sum_{n=m+1}^{K}\left\|\omega_m-\omega_n\right\|_{\mathcal{F}}^2+C\sum_{i=1}^{\ell}\sum_{m\neq y_i}\xi_i^{(m)}\,, \tag{8}$$

is solved subject to

$$\left\langle\omega_{y_i}-\omega_m,\phi\left(\mathbf{x}_i\right)\right\rangle_{\mathcal{F}}\quad\geq\quad 1-\xi_i^{(m)}\,, \tag{9}$$
$$\xi_i^{(m)}\quad\geq\quad 0\,,\quad i=1,\ldots,\ell\quad m\in\mathcal{Y}\backslash y_i\,. \tag{10}$$

Wolfe's dual optimization problem is obtained by introducing Lagrange multipliers, so the hyperplane separating classes θ_r and θ_s can be expressed as

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 161-166

$$h(\mathbf{x}) =$$

$$\frac{1}{K}\left(\sum_{i:y_i=r} A_i k(\mathbf{x}_i,\mathbf{x}) - \sum_{i:y_i=s} A_i k(\mathbf{x}_i,\mathbf{x}) - \sum_{i=1}^{\ell}\left(\alpha_i^{(r)} - \alpha_i^{(s)}\right) k(\mathbf{x}_i,\mathbf{x})\right). \quad (11)$$

2.3 Combining RLP with SVMC

[2] derives an hybrid machine combining a linear programming method, RLP, for the multi-class problem, k-RLP, and the QP SVM method. The novel multi-class classification algorithm, M-SVM, builds piecewise-linear discriminants.

In this case, original notation is very different from the subscripts' notation in former subsections. Details of this transformation are omitted due to the lack of space. The associated QP to be solved can be written as $\arg\min R_{M-SVM}(\omega,\xi)$

$$= \frac{1}{2}\left[\sum_{m=1}^{K}\|\omega_m\|_{\mathcal{F}}^2 + \sum_{m=1}^{K-1}\sum_{n=m+1}^{K}\|\omega_m-\omega_n\|_{\mathcal{F}}^2\right] + C\sum_{i=1}^{\ell}\sum_{m\neq y_i}\xi_i^{(m)} \quad (12)$$

subject to (9)–(10), and the complete decision function is identical to (7).

2.4 The Relation with the Bayes Rule

In [6] the SVM paradigm is treated from the statistical point of view and the Bayes optimal classification rule is applied.

A new notation must be treated, however after some translations the optimization problem for the standard case can de defined as

$$\arg\min R_{Bayes}(\omega,\xi) = \frac{1}{2}\sum_{m=1}^{K}\|\omega_m\|_{\mathcal{F}}^2 + C\sum_{i=1}^{\ell}\sum_{m\neq y_i}\xi_i^{(m)}, \quad (13)$$

subject to

$$\langle \omega_{y_i} - \omega_m, \phi(\mathbf{x}_i)\rangle_{\mathcal{F}} \geq \tfrac{K}{K-1} - \xi_i^{(m)}, \quad (14)$$

$$\xi_i^{(m)} \geq 0, \quad i=1,\ldots,\ell \quad m\in\mathcal{Y}\backslash y_i. \quad (15)$$

Treating the primal QP problem in a similar way as (1), dual formulation is

$$W_{Bayes}(\alpha) = \tfrac{K}{K-1}\sum_{i=1}^{\ell}\sum_{m=1}^{K}\alpha_i^{(m)} + \\ + \sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\sum_{m=1}^{K}\left(-\frac{1}{2}c_j^{(y_i)}A_iA_j + \alpha_i^{(m)}\alpha_j^{(y_i)} - \frac{1}{2}\alpha_i^{(m)}\alpha_j^{(m)}\right)\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle_{\mathcal{F}}, \quad (16)$$

subject to $0\leq\alpha_i^{(m)}\leq C$, $\alpha_i^{(y_i)} = 0$, $i=1,\ldots,\ell$ $m\in\mathcal{Y}\backslash y_i$.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 161-166

3 The New General Formulation

It can be observed that the third method, expressed in a useful notation, has in its risk functional (12) a conjunction of the former ones, (1) and (8), and it is a good starting point to develop a more general framework.

A unified multi-class risk functional is defined with the associated QP problem $\arg\min R_{Uni}(\omega, \xi)$

$$= A \sum_{m=1}^{K} \|\omega_m\|_{\mathcal{F}}^2 + B \sum_{m=1}^{K-1} \sum_{n=m+1}^{K} \|\omega_m - \omega_n\|_{\mathcal{F}}^2 + C \sum_{i=1}^{\ell} \sum_{m \neq y_i} \xi_i^{(m)} \qquad (17)$$

subject to

$$\langle \omega_{y_i} - \omega_m, \phi(\mathbf{x}_i) \rangle_{\mathcal{F}} \geq D - \xi_i^{(m)}, \qquad (18)$$

$$\xi_i^{(m)} \geq 0, \quad i = 1, \ldots, \ell \quad m \in \mathcal{Y} \backslash y_i. \qquad (19)$$

All of the four former algorithms can be displayed with this new formulation[4]

$$\begin{aligned} R_{Ext}(\omega, \xi) &= R_{Uni}(\omega, \xi)_{A=\frac{1}{2}, B=0, D=2} \\ R_{Conv}(\omega, \xi) &= R_{Uni}(\omega, \xi)_{A=0, B=\frac{1}{2}, D=1} \\ R_{M-SVM}(\omega, \xi) &= R_{Uni}(\omega, \xi)_{A=B=\frac{1}{2}, D=1} \\ R_{Bayes}(\omega, \xi) &= R_{Uni}(\omega, \xi)_{A=\frac{1}{2}, B=0, D=\frac{K}{K-1}} \end{aligned} \qquad (20)$$

and constraints are similar in all the cases.

3.1 A First Theoretical Result

Comparisons between algorithms are easier with the new framework. So, the Wolfe's dual optimization problem for $R_{Uni}(\omega, \xi)_{A=\frac{1}{2}, B=0, D=1}$ is (6), while the Wolfe's dual for $R_{Uni}(\omega, \xi)_{A=0, B=\frac{1}{2}, D=1}$ is $W(\alpha) =$

$$\frac{1}{2K} \sum_{i,j=1}^{\ell} \sum_{m=1}^{K} \left[2\alpha_i^{(m)} \alpha_j^{(y_i)} - c_j^{(y_i)} A_i A_j - \alpha_i^{(m)} \alpha_j^{(m)} \right] \cdot k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{\ell} A_i, \qquad (21)$$

subject to $0 \leq \alpha_i^{(m)} \leq C, \quad \alpha_i^{(y_i)} = 0 \quad i = 1, \ldots, \ell \quad m \in \mathcal{Y} \backslash y_i$.

By observing the dual formulation in (6) and (21), it can be conclude that the main difference is the term $\frac{1}{K}$ in the second dual risk formulation, although the restrictions are equivalent.

This new tool makes the analysis of the performance and behaviour of these approaches easier. An example illustrating this relation follows: if the very usual gaussian kernel

$$k(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{\sigma^2} \right), \qquad (22)$$

[4] $R_{Vap}(\omega, \xi) = R_{Uni}(\omega, \xi)_{A=\frac{1}{2}, B=0, D=1}$

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 161-166

is choosen in (6), being the width σ, it could be possible to obtain the equivalence between (6) and (21) if an artificial width $\widetilde{\sigma}$ is defined in the kernel of (21) in the form

$$\widetilde{\sigma}^2 = \sigma^2 \cdot \frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{\|\mathbf{x} - \mathbf{x}_i\|^2 - \sigma^2 \ln K} > \sigma^2 \,. \tag{23}$$

This observation leads to conclude that the solutions for the second case have a more generalized behaviour than the previous one when gaussian kernels are used because the shape of the gaussian functions are wider.

4 Conclusions and Further Work

Several extensions have been recently derived to deal multi-category problems by using 'all data at once' SVMs. A unified algorithmic framework has been created for these approaches extending the bi-class SVM learning machine for multi-class purposes. With the new environment, we easily obtain a first theoretical result about generalization using gaussian kernels in an easy form, showing that this is a powerful tool to drive this extension. Work in this open research area and relations with other extensions and studies from different authors will be continued.

References

[1] Angulo, C., Català, A.: K-SVCR. A Multi-class Support Vector Machine. Proc. European Conference on Machine Learning ECML'00 (2000)

[2] Bredensteiner, E.J., Bennett, K.P.: Multicategory classification by support vector machines. Computational Optimizations and Applications. **12**(1999)

[3] Dietterich, T.G., Bakiri G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research. **2** (1995)

[4] Guermeur Y., Elisseeff A., Paugam-Moisy, H.: A new multi-class SVM based on a uniform convergence result. Proceedings of the Internatinal Joint Conference on Neural Networks IJCNN'00 (2000)

[5] Kressel, U.: Pairwise classification and support vector machines. In Schölkopf, B., Burges C.J.C., Smola, A.J. (eds.): Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA (1999)

[6] Lee, Y., Lin Y., Wahba, G: Multicategory Support Vector Machines. Tech Report No. 1043. University of Wisconsin, USA (2001)

[7] Vapnik, V.: Statistical learning theory. Wiley, New York (1998)

[8] Weston J., Watkins, C.: Multi-class support vector machines. CSD-TR-98-04. Royal Holloway, University of London. Egham, UK (1998)