Neural Networks 23 (2010) 607–613

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A neural support vector machine

Magnus Jändel *

Agora for Biosystems, Box 57 SE-193 22, Sigtuna, Sweden
Swedish Defence Research Agency, SE-164 90, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 24 November 2008
Received in revised form 4 October 2009
Accepted 2 January 2010

Keywords:
Support vector machine
Neural systems
Pattern recognition
Perceptual learning
Associative memory
Olfactory system

ABSTRACT

Support vector machines are state-of-the-art pattern recognition algorithms that are well founded in optimization and generalization theory but not obviously applicable to the brain. This paper presents Bio-SVM, a biologically feasible support vector machine. An unstable associative memory oscillates between support vectors and interacts with a feed-forward classification pathway. Kernel neurons blend support vectors and sensory input. Downstream temporal integration generates the classification. Instant learning of surprising events and off-line tuning of support vector weights trains the system. Emotion-based learning, forgetting trivia, sleep and brain oscillations are phenomena that agree with the Bio-SVM model. A mapping to the olfactory system is suggested.

1. Introduction

Human and animal brains excel in complex classifications. Friend or foe? Edible or poisonous? Survival depends on such quick appraisals. How does the brain implement trainable general-purpose classifiers that learn instantly, yet avoid overwriting relevant lessons and match outputs to appropriate behaviours?

Instant learning is vital in an unforgiving environment. Overwriting or diluting old but still valid experiences is dangerous. Yet there is not enough memory or search resources for remembering everything. Memory must be managed so that vital knowledge is conserved while trivial experiences are discarded. Connecting the output of plastic neural classifiers to predetermined behavioural triggers is crucial. Predator scent detection must for example be coupled to flight behaviour. Haberly (2001) found, however, that biologically plausible algorithms for trainable pattern recognition generating predetermined output codes are in short supply.

This paper introduces Bio-SVM, a biologically feasible support vector machine that instantly learns surprising examples, forgets trivial examples and trains an optimal generalizing classifier with predetermined output codes. Bio-SVM is consistent with the observed mix of fast and slow brain oscillations and maps well to the architecture of the olfactory system.

The generic pattern recognition task is to classify a test sample by generalizing from known classifications of training examples. We shall only consider binary classifications. Multi-value classifications can readily be produced by a bank of binary classifiers. The training set S consists of examples $(\mathbf{x}, y)$, where $\mathbf{x}$ is a real-valued input vector and $y \in \{+1, -1\}$ indicates the correct classification. Bold letters signify vector quantities. The training examples are presented in a batch or more realistically one-by-one in online learning.

Support vector machines (SVMs) (see Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002 for reviews) have recently emerged as a strong alternative for any classification application. An SVM works by projecting input vectors $\mathbf{x}$ to a high-dimensional feature space. Features $\boldsymbol{\phi}(\mathbf{x})$ are typically non-linear functions of the input vector. The training algorithm finds a hyperplane in feature space separating positive cases from negative cases with maximum margin. The set of feature-space hyperplanes provides the broad hypothesis domain that is vital for solving substantial classification tasks. Enforcing maximal margins ensures a generalization performance that is optimal in a certain well-defined sense. The key insight of SVM pioneers Boser, Guyon, and Vapnik (1992) is that the SVM optimization problem can be solved without explicitly constructing the feature space.

The solution to a classification problem is the set of support vectors **SV**. Each support vector $\mathbf{x}_i$ is drawn from the training examples and has an associated positive real-valued weight α_i. The support vectors are borderline members of the training data used for defining the partitioning feature-space hyperplane. Positive support vectors are close to the negative domain. Negative support vectors are similarly bordering to the positive realm. An SVM classifies test samples $\mathbf{x}$ using a real-valued classification function $f(\mathbf{x})$. The test sample belongs to the negative class $y = -1$ if $f(\mathbf{x}) < 0$ and to the

* Corresponding address: Vårvägen 10, SE-19460 Upplands Väsby, Sweden. Tel.: +46 709277264; fax: +46 855503700.
E-mail addresses: magnus@jaendel.se, magjan@foi.se.

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.01.002

positive class $y = 1$ otherwise. The classification function is

$$f(\mathbf{x}) = \sum_{i \in \mathbf{SV}} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b, \tag{1}$$

where b is a bias parameter. The positive definite kernel function K defines the implicit projection to feature space. For a given pair of input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, $K(\mathbf{x}_i, \mathbf{x}_j)$ is a measure of alignment in feature space.

2. The Bio-SVM model

SVMs are correctly viewed as founded on rigorous mathematics rather than biological analogies. Solution algorithms suggest implementation in a digital computer. There is, however, one aspect of SVMs that stands out as similar to biological systems. An SVM ignores typical examples but pays attention to borderline cases and outliers. It remembers surprises and forgets run-of-the-mill events. Life learns also from odd emotionally charged events. We remember the support vectors. Given their mathematical soundness, efficiency and a certain high-level similarity to biological learning, could SVMs be implemented in the brain?

This section casts the abstract SVM concept into a form that can be implemented by biological neural systems—the Bio-SVM model. This hypothesis is then compared to the gross features of brain pattern recognition systems.

2.1. Zero-bias ν-SVM

We must first find a formal SVM model that is malleable to neural form. The base-line is ν-SVM (Schölkopf, Smola, Williamson, & Bartlett, 2000), a soft-margin SVM in which a dimensionless parameter $0 < \nu < 1$ controls the trade-off between generalization and accuracy. Soft margin means that outlier support vectors may violate margins. Such mavericks are expected in noisy training sets. Schölkopf et al. (2000) show that ν is an upper bound on the fraction of margin errors. The ν-SVM model is solved, for a set of m training examples, by maximizing the dual objective function,

$$W(\boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

subject to

$$0 \le \alpha_i \le \frac{1}{m} \tag{3}$$

and

$$\sum_{i=1}^{m} \alpha_i \ge \nu. \tag{4}$$

The classification function is defined by Eq. (1). The solution to this problem is the optimal feature-space hyperplane.

We use a modified version of ν-SVM in which the bias parameter in Eq. (1) is set to zero. This is achieved by embedding the feature space vector $\boldsymbol{\phi}(\mathbf{x})$ of the original problem in a larger space $\{\boldsymbol{\phi}(\mathbf{x}), \tau\}$, thus increasing the dimensionality by one (Cristianini & Shawe-Taylor, 2000). This operation corresponds to replacing the old kernel K with a new kernel $K + \tau^2$. Removing the bias means less freedom for optimization and thus potentially smaller feature space margin, leading to reduced generalization performance (see Cristianini & Shawe-Taylor, 2000, p. 131). As explained in Section 2.5, it is, however, an essential simplification for mapping the model to a biological substrate.

The solution of the ν-SVM problem in Eqs. (2)–(4) is the weight vector $\boldsymbol{\alpha}$. We need a solution algorithm that is suitable for physiological modelling even if it may be suboptimal as a serial computer algorithm. We note that in general there exists an optimal solution in the α-space hyperplane,

$$\sum_{i=1}^{m} \alpha_i = \nu, \tag{5}$$

(Chang & Lin, 2001). The strategy is to start at an arbitrary point in the allowed domain of the α-hyperplane, e.g. by initializing all α_i to ν/m, and then follow the projection of the gradient of $W(\boldsymbol{\alpha})$ in the α-hyperplane until an optimum is found.

The gradient projection is

$$\mathbf{grad_P}(W) = \mathbf{grad}(W) - \mathbf{e}_\perp(\mathbf{grad}(W) \cdot \mathbf{e}_\perp), \tag{6}$$

where $\mathbf{e}_\perp$is the unit normal vector of the α-hyperplane and $\mathbf{grad}(W) = (\frac{\partial W}{\partial \alpha_1}, \frac{\partial W}{\partial \alpha_2}, \ldots, \frac{\partial W}{\partial \alpha_m})$. The ith gradient component is

$$\frac{\partial W}{\partial \alpha_i} = -y_i \sum_{j=1}^{m} y_j \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) = -y_i f(\mathbf{x}_i) = -marg_i, \tag{7}$$

where $marg_i$ is the margin in feature space between the example and the classification hyperplane. A positive margin means that the example is classified correctly. The ith component of the gradient projection is, therefore,

$$\mathbf{grad_P}(W)_i = \langle marg \rangle - marg_i, \tag{8}$$

where $\langle marg \rangle = \frac{1}{m} \sum_{j=1}^{m} y_j f(\mathbf{x}_j)$ is the average margin. Each weight shall hence be updated in proportion to the difference between the average margin and the margin of the associated example. This rule strives to make the margin of each example equal to the average margin as the hypothesis $\boldsymbol{\alpha}$ progresses towards the optimum. It is not always possible to reach equality. The converged ν-SVM partitions the training examples into three distinct sets.

Trivial examples are non-support vectors. Their weights are driven to zero since the margin of such examples is larger than the average margin. Note that trivial examples can be removed from the training set once a solution has been found.

Outliers are possibly misclassified support vectors that consistently fall beyond of the average margin. Their weights are pushed to the maximum value $1/m$.

Regular support vectors converge to the average margin. Their weights fall within $0 < \alpha_i < 1/m$.

Eq. (2) is quadratic with respect to $\boldsymbol{\alpha}$ and the maximum is sought in the $\boldsymbol{\alpha}$-space hyperplane defined by Eqs. (3) and (5). This guarantees that there are no false maxima (see Chang & Lin, 2001 for a proof). It is hence easy to evaluate convergence. With plenty of time and computational resources one can simply move in very small steps along $\mathbf{grad_P}(W)$ until the maximum is found. The challenge is to find a biological apparatus that does just this.

2.2. The Bio-SVM concept

The general architecture and key operational processes for mapping zero-bias ν-SVM to brain systems are first outlined here and then detailed in the following sections. The main modules of the Bio-SVM are the Oscillating Memory (OM) for learning and storing support vectors and the Classification Pathway (CP) for performing classifications. The OM is the only plastic part of the system. The overall architecture is shown in Fig. 1.

The Bio-SVM executes three processes:

(1) **Classification**, where sensory inputs are classified.
(2) **Surprise learning**, where new training examples are engraved. A supervising unit, called the Critic, detects failed classifications and triggers the OM to remember the anomalous event.
(3) **Importance learning**, where trivial examples are forgotten and support vectors get optimal weights. The OM inputs training examples to the CP while the brain sleeps and adjusts weights according to resulting feedback. Examples are forgotten if weights consistently fall to zero.

Higher brain systems control which process to employ.