ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 257-262

Support Vector Machines vs Multi-Layer Perceptron in Particle Identification

N.Barabino[1], M.Pallavicini[2], A.Petrolini[1,2], M.Pontil[3,1], A.Verri[4,3]

[1] DIFI, Università di Genova (I)

[2]INFN Sezione di Genova (I)

[3]Center for Biological and Computational Learning, MIT
Cambridge (US)

[4] INFM - DISI, Università di Genova (I)

Abstract. In this paper we evaluate the performance of Support Vector Machines (SVMs) and Multi-Layer Perceptrons (MLPs) on two different problems of Particle Identification in High Energy Physics experiments. The obtained results indicate that SVMs and MLPs tend to perform very similarly.

1. Introduction

Support Vector Machines (SVMs) have been recently introduced as a technique for pattern recognition which approximately implements *structural risk* minimization. Whereas previous techniques, like Multi-Layer Perceptrons (MLPs), are based on the minimization of the *empirical risk*, that is the minimization of the number of misclassified points of the training set, SVMs minimize a functional which is the sum of two terms. The first term is the empirical risk, the second term controls the confidence with which the obtained separating surface behaves on previously unseen data points. SVMs are attracting increasing attention because they rely on a solid statistical foundation [8,9] and appear to perform quite effectively in many different applications [2-5,7]. A clear advantage of SVMs over MLPs is due to the intuitive interpretation and understanding of the behavior of SVMs on each particular problem. After training, the separating surface is expressed as a certain linear combination of a given kernel function centered at some of the data points (named *support vectors*). All the remaining points of the training set are effectively discarded and the classification of new points is obtained solely in terms of the support vectors.

In this paper we compare SVMs against MLPs on two different problems of Particle Identification (PI) in High Energy Physics (HEP) experiments. MLPs are commonly accepted by the HEP community as a standard technique for

PI. The aim of our work is to assess the potential of SVMs as an alternative (or complementary) method for solving pattern recognition problems, with emphasis on PI problems. We consider two PI problems. The first uses simulated data from DELPHI about the process $e^+e^- \longrightarrow \tau^+\tau^-$, the second real data about the process $J/\psi \longrightarrow e^-e^+$. The input points are vectors in 14- and 8-dimensional spaces respectively. The training and test sets vary from a few to several thousands of points.

The organization of the paper is as follows. In section 2 we review briefly the main properties of SVMs for pattern recognition. In section 3 we present a short description of the physical origin of the used data. The experiments performed with SVMs and MLPs are discussed in section 4. Finally, we discuss the conclusions which can be drawn from our analysis in section 5.

2. Support Vector Machines

Let us briefly review the theory of SVMs. For a more detailed account and the connection between SVMs and the minimization of the structural risk we refer to [9].

We assume we are given a set S of N points $\mathbf{x}_i \in \mathbb{R}^n$ $(i = 1, 2, \ldots, N)$. Each point $\mathbf{x}_i$ belongs to either of two classes identified by the label $y_i \in \{-1, 1\}$. In the further assmption that the two classes can be linearly separable, the goal is to establish the hyperplane, named Optimal Separating Hyperplane 9OSH), that divides S leaving all the points of the same class on the same side while maximizing the distance of the closest point. It can be shown [9] that the OSH is the solution to the problem

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}\mathbf{w}\cdot\mathbf{w} \\ \text{subject to} & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N \end{array}$$

where $\mathbf{w}$ is the normal of the hyperplane, and b/w the distance of the hyperplane from the origin. If we denote with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ the N nonnegative Lagrange multipliers associated with the constraints, the solution to this problem is equivalent to determining the solution to the *dual* problem

$$\begin{array}{ll} \text{Maximize} & -\frac{1}{2}\boldsymbol{\alpha}^\top D\boldsymbol{\alpha} + \sum \alpha_i \\ \text{subject to} & \sum y_i\alpha_i = 0 \\ & \boldsymbol{\alpha} \geq 0, \end{array}$$

where the sums are for $i = 1, 2, \ldots, N$, and D is an $N \times N$ matrix such that

$$D_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j. \tag{1}$$

The solution for $\bar{\mathbf{w}}$ reads

$$\bar{\mathbf{w}} = \sum_{i=1}^{N} \bar{\alpha}_i y_i \mathbf{x}_i, \tag{2}$$

The only $\bar{\alpha}_i$ that can be nonzero in Eq.(2) are those for which the constraints of the first problem are satisfied with the equality sign. Since most of the $\bar{\alpha}_i$

are usually null, the vector $\bar{\mathbf{w}}$ is a linear combination of a often relatively small percentage of the points $\mathbf{x}_i$. These points are termed *support vectors* because they are the only points of S needed to determine the *OSH*.

The problem of classifying a new data point $\mathbf{x}$ is now simply solved by looking at the sign of

$$\bar{\mathbf{w}} \cdot \mathbf{x} + \bar{b}$$

with $\bar{b}$ obtained from the Khun Tucker conditions.

If the set S cannot be separated by a hyperplane, the previous analysis can be generalized by introducing N nonnegative variables $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N)$ such that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, N. \tag{3}$$

The so called generalized *OSH* is then regarded as the solution to

$$\begin{array}{ll} \text{Maximize} & \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C \sum \xi_i \\ \text{subject to} & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, 2, \ldots, N \\ & \boldsymbol{\xi} \geq 0. \end{array}$$

Similarly to the linearly separable case, the dual formulation requires the solution of a quadratic programming problem with linear constraints. Once again it turns out that the points that satisfy the constraints of above with the equality sign are termed support vectors and are the only points needed to determine the decision surface.

The entire construction can also be extended rather naturally to include nonlinear separating surfaces [9]. Each point $\mathbf{x}$ in input space is mapped into a point $\mathbf{z} = \phi(\mathbf{x})$ of a higher dimensional feature space (possibly of infinite dimension). The mapping ϕ is subject to the condition that the dot product $< \phi(\mathbf{x}), \phi(\mathbf{y}) >$ in feature space can be rewritten through a kernel function $K = K(\mathbf{x}, \mathbf{y})$. Admissible kernel functions, for example, are the polynomial kernel of n-th degree

$$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^n - 1$$

or the Gaussian kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(||\mathbf{x} - \mathbf{y}||/2\sigma^2\right)$$

3. Data description

In this paper we consider data from two different HEP experiments. In the first we use simulated data, while in the second data gathered from a real experiment.

The simulated data of the present analysis are the same data which have been used in [6]. Event simulation is widely used to define methods of extraction of signal from background, study the effects of possible systematic errors, and model the detector effects on the events. Here, events where a $\tau^+\tau^-$ pair is produced in the decay of a Z^0 from the annihilation of a e^+e^- pair must be selected from the overwhelming background of Z^0 decays into hadrons. The simulated events, obtained according to a procedure described in [6] are in the same format as the real data. A number of physical variables pointing out certain characteristics of every kind of event, 14 in the present case, are then defined and the separation between different classes of events is obtained in terms of the different probability distribution of the variables for the different classes.

The real data are taken from experiment E760 at Fermilab [1]. This experiment was searching for events with inclusive J/ψ decaying to e^+e^- produced in $p-\bar{p}$ interactions. Each track is described as a feature point in an 8-dimensional space. As *electrons* for the training set we used $p-\bar{p} \longrightarrow \chi_2 \longrightarrow J/\psi\gamma \longrightarrow e^+e^-\gamma$ events selected requiring good exclusive kinematic fit (prob. > 0.2)[1]. The *background* for the training set was selected using 2pb^{-1} data taken at energy far away from charmonium resonances ($E_{cm} = 3510$MeV), keeping all events with invariant mass above 2.4GeV/c^2. The goal was to identify background events that almost perfectly simulate the signal. We ended up with a training set with far less than 1% of contaminated data (that is, electrons in background or background in electrons data).

4. Performed experiments

For MLPs we use JETNET3.0, a package for the training of neural networks developed for PI applications and available via ftp from `ftp://www.then.lu.se`. The best results have been obtained with a number of hidden units equal to twice the number of the input variables (that is, 28 and 16 respectively).

For SVMs we use the implementation available in our lab, based on the decomposition algorithm proposed in [4] and a solver which transforms the dual problem, a QP problem with linear constraints, in a linearly complementary problem. We used polynomial and Gaussian kernels for the two experiments respectively. The degree of the polynomial and the variance of the Guassian kernel (over ten possible values) were determined by looking at the minimum of the quantity R^2w^2 on the training set, with R the radius of the minimal ball enclosing the training d ata in feature space and w the computed margin.

The overall results for the two experiments are summarised in Tables 1 and 2. The reported results refer to the percentage of correct classifications obtained on the entire test set using all the points of the training set. For the case of simulated data the training and test sets consist of 10000 and 5000 points, while for the case of real data of 3600 and 1400 respectively.

As it can easily be inferred from Tables 1 and 2, the two methods give almost indistiguinshable results. The SVMs give slightly better recognition rates on both simulated and real data but probably not enough to be statistically significant (though in the case of simulated data SVMs performed consistently better than MLPs).

[1] These selection criteria provide a very clean set of electron tracks without introducing bias in the feature points.

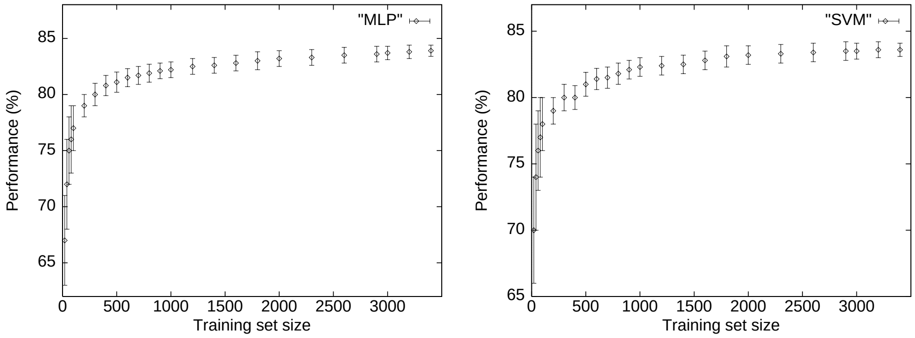
Table 1: Overall results on simulated data

Method	Error rate
Best MLP (Manhattan alg.)	95.5
Polynomial SVM (8th degree with C = 10)	96.6

Table 2: Overall results on real data

Method	Error rate
Best MLP (Manhattan alg.)	83.6
Gaussian kernel SVM (σ = 1.5 with C = 10)	84.4

Extensive experimentation indicate that the two methods tend to give similar results also on training sets of reduced size. This can be seen, for example, in Figure 4. which shows the performance of MLP and SVM in the case of real data as a function of the training set size. For both methods, the error bars mark the best and worst performance obtained by random sampling of the training sets.



5. Conclusions

In this paper we have presented results of a comparison between MLPs and SVMs on two problems of Particle Identification. The results obtained so far indicate that the methods tend to produce very similar results. In the two specific problems we have considered SVM perform always at least as well as MLP (that is, within the error margin of the performed experiments) and, in the case of simulated data, consistently better.

While more experiments are needed to reach a final virdict, we can already draw a number of conclusions. First, SVMs seem to work well even in the presence of large training sets drawn from input spaces of relatively small dimension. Second, contrary to theoretical expectations, we have found an unusually large number of support vectors (often more than 50% of the points of the training set). This is probably due to the large amount of noise affecting the input points. Third, and finally, the procedure of selecting the appropriate degree for a polynomial kernel or the variance for the Gaussian kernel by means of the empirical minimization of the quantity R^2w^2 proved to be rather effective.

References

[1] T. Armstrong et al., *Phys. Rev. Lett.* **69**, 2337 (1992).

[2] A. Ganapathiraju, J. Hamaker, and J. Picone, Support Vector Machines for Speech Recognition, in *Proc. Int. Conf. on Spoken Language Processing*, Sidney (1998).

[3] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, A General Framework for Object Detection, in *Proc. CVPR*, Puertorico (1997).

[4] E. Osuna, R. Freund, and F. Girosi, Training Support Vector Machines: an Application to Face Detection, in *Proc. CVPR*, Puertorico (1997).

[5] C. Papageorgiou, M. Oren, and T. Poggio, A General Framework for Object Detection, in *Proc. ICCV*, Bombay (1998).

[6] A. Petrolini, The Use of Neural Network Classifiers for Higgs Searches, *Int. J. Mod. Phys* **C3**, 611 (1992).

[7] M. Pontil and A. Verri, Object Recognition with Support Vector Machines, *IEEE Trans. on PAMI* **20**, 637–646 (1998).

[8] V. Vapnik, **The Nature of Statistical Learning Theory**, Springer (1995).

[9] V. Vapnik, **Statistical Learning Theory**, Wiley (1998).