Prediction of Probability of Survival in Critically Ill Patients Optimizing the Area Under the ROC Curve*

Oscar Luaces⋆**, José R. Quevedo**⋆**, Francisco Taboada**†**,**
Guillermo M. Albaiceta†**, Antonio Bahamonde**⋆

⋆Artificial Intelligence Center
University of Oviedo at Gijón
Asturias - Spain

†Hospital Univ. Central de Asturias (HUCA)
University of Oviedo
Asturias - Spain

Abstract

The paper presents a support vector method for estimating probabilities in a real world problem: the prediction of probability of survival in critically ill patients. The standard procedure with Support Vectors Machines uses Platt's method to fit a sigmoid that transforms continuous outputs into probabilities. The method proposed here exploits the difference between maximizing the AUC and minimizing the error rate in binary classification tasks. The conclusion is that it is preferable to optimize the AUC first (using a multivariate SVM) to then fit a sigmoid. We provide experimental evidence in favor of our proposal. For this purpose, we used data collected in general ICUs at 10 hospitals in Spain; 6 of these include coronary patients, while the other 4 do not treat coronary diseases. The total number of patients considered in our study was 2501.

1 Introduction

The available models for predicting outcomes in intensive care units (ICU) are usually scoring systems that estimate the probability of hospital mortality of critically ill adults. This is the case of APACHE (*Acute Physiology And Chronic Health Evaluation*) [Knaus *et al.*, 1991], SAPS (*Simplified Acute Physiology Score*) [Le Gall *et al.*, 1984], and MPM (*Mortality Probability Models*) [Lemeshow *et al.*, 1993]. The score functions of these predictors were induced from data on thousands of patients using logistic regression. The data required by these systems come from monitoring devices, clinical analysis, and demographic and diagnostic features of patients. So, APACHE III includes age, 16 acute physiologic variables that use the worst value from the first 24 hours in the ICU (temperature, heart rate, blood pressure, respiratory rate, oxygenation, acid-base status, serum sodium, serum blood urea nitrogen, serum creatinine, serum albumin, serum bilirrubin, serum glucose, white cell count, hematocrit, itemized Glasgow Coma Scale score, and urine output), preexisting functional limitations, major comorbidities, and treatment location immediately prior to ICU admission.

These prognostic models are mainly used to measure the efficiency of ICU treatments. The risk stratification of patients allows comparison of the observed outcomes versus accepted standards provided by score functions. ICU assessment is very important since it is estimated that end-of-life care consumes 10% to 12% of all healthcare costs. Moreover, in 2001 the average daily cost per patient in ICUs was about $3000 in the USA [Provonost and Angus, 2001]. On the other hand, the literature also shows that prognoses have constituted an important dimension of critical care, as patients and their families seek predictions about the duration and outcome of illness [Lemeshow *et al.*, 1993].

In this paper we propose a new method for learning probabilities that will be tested on the probabilities of survival in ICU patients. The method makes intensive use of the so-called Support Vector Machines (SVM), a powerful family of algorithms for learning classification and regression tasks. When used for binary classification, SVM learn hypotheses that return continuous numbers: positive values for cases of one class, and negative for the other class.

On the other hand, to measure the performance of predictions in medicine, and in general when classes are very unbalanced, the misclassification rate (or accuracy) is usually inadequate. Frequently, the Area Under a *receiver operating characteristic* (ROC) Curve (AUC for short) is used. This amount can be interpreted as the degree of coherence between a continuous output (such as the probability, or the continuous output of an SVM) and a binary classification. It is important to emphasize that that coherence is established in terms of orderings. For this purpose, continuous outputs or scores are used to rank available cases, while classes in the ICU problem are codified by '$+1$' when the patient has survived, and '-1' otherwise.

In this context, Hanley and McNeil [1982] showed that the AUC is the probability of a correct ranking; in other words, it is the probability that a randomly chosen subject of class '$+1$' is (correctly) ranked with greater output than a randomly chosen subject of class '-1'. Therefore, AUC coincides with the value of the Wilcoxon-Mann-Whitney statistic.

Additionally, there are other measures of the goodness of probability estimations; for instance, the Brier score is the average of quadratic deviations of true and predicted probabili-

*The research reported here is supported in part under grant TIN2005-08288 from the MEC (Ministerio de Educación y Ciencia of Spain). The authors acknowledge the work of the Grecia Group in the collection of data.

ties. The relation between AUC and Brier scores was studied in [Ikeda *et al.*, 2002]. However, the relationship found is guaranteed only under very restrictive conditions that are difficult to check in real world cases. Moreover, the relationship is not always direct even in the case study reported in the experimental section of this paper.

To learn a probability distribution using SVM, it is crucial to transform their scores or continuous outputs into probabilities. But this is what a method presented by Platt [2000] does. The core idea is to fit a sigmoid, using a maximum likelihood procedure. The novelty of the proposal reported in this paper is that we postulate that to compute Platt's sigmoid it is better to look for an optimum AUC first than to minimize the error rate with a classification SVM. For this reason, in Section 2.3 we shall discuss how to optimize the AUC with a Support Vector method [Herbrich *et al.*, 2000; Joachims, 2005].

The rationale behind our proposal is that the quality of the sigmoid fit depends on the quality of the ranking of the scores. If most of the cases with a higher score than a given one of class y have a class greater than y, then the task of the sigmoid can be easily accomplished, and the performance of the final probability is nearly optimal.

At the end of the paper we provide experimental evidence in favor of our proposal, comparing it with other alternative approaches. For this purpose, we used data collected in general ICUs at 10 hospitals in Spain, 6 of which include coronary patients, while the other 4 do not treat coronary diseases. The total number of patients considered in our study was 2501, 19.83% of whom did not survive.

2 Predicting probabilities

In this section we shall start off by reviewing a standard method for learning probabilities based on Support Vector Machines to then present our proposal. But first of all we must realize that the performance of classification learners is not satisfactory in the ICU problem; otherwise, nobody would turn to probabilities. This is a general situation in medicine, as well as in other fields; accurate crisp predictions are difficult to make, but some useful knowledge can be drawn from data.

The section will end with the description of a straightforward approach for learning probabilities using regression. This method will be used as a baseline for measuring the merits of the other options.

2.1 The goodness of probability predictions

Let $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ be a training set for a learning task in which a function (or hypothesis) is sought that is able to return outputs y_i from entries $\boldsymbol{x}_i$ of an input space $\mathcal{X}$. An important issue when we are learning is to fix the way in which we are going to measure the quality of the result. In fact, given S formally, the aim of learning is to find a hypothesis h (from a given hypothesis space) that minimizes the average *loss* extended over the set of independently identically distributed (i.i.d.) test sets S', usually represented by $\Delta(h, S')$.

In the ICU problem, training and test examples have no probability attached, they are labeled with $+1$ or -1. Therefore, we shall assume that the *true* class probability, $Pr^{true}(y = +1|\boldsymbol{x})$, is 1 when the class of $\boldsymbol{x}$, y, is $+1$ and 0 otherwise.

In general, when predictions are discrete probability distributions, there is basically one standard loss function: the average quadratic deviation. If there are two possible outputs, the probability loss is given by

$$\Delta_{\mathrm{Pr}}(h, S') = \frac{1}{|S'|} \sum_{\boldsymbol{x}'_i \in S'} (h(\boldsymbol{x}'_i) - p_i)^2 \qquad (1)$$

where the hypothesis h returns the estimation of the probability $h(\boldsymbol{x}) = Pr(y = +1|\boldsymbol{x})$, and p_i stands for the observed probability of the i-th case, $p_i = Pr^{true}(y = +1|\boldsymbol{x}_i)$.

The measurement in Equation (1) is frequently used in medicine and meteorology, and is known as the *Brier* [1950] index or *score*. If the number of possible outputs is greater than two, the estimated probabilities can be seen as a vector, and the Mean Square of the Euclidean (MSE) distance from predicted and observed probabilities is then used; see, for instance [Melville *et al.*, 2005]. It can be seen that, in the ICU problem, MSE is 2 times the Brier score.

2.2 Optimizing accuracy plus a sigmoidal transformation

The straightforward approach to the ICU problem is a binary classification SVM followed by a sigmoid estimated using Platt's method [2000]. Thus, given the training set S, we can use a transformation ϕ defined from input entries in $\mathcal{X}$ into a feature space $\mathcal{H}$, where classes should be mostly separable by means of a linear function. As is well known, $\mathcal{H}$ must have an inner product $\langle , \rangle$, and

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \qquad (2)$$

is called the kernel function of the transformation. We shall use the rbf kernel that is defined by

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \qquad (3)$$

The work of the SVM consists in solving the following convex optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w},\xi} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{n} \xi_i, \qquad (4) \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i) \rangle + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, n \end{aligned}$$

Then, the classification is accomplished by the hypothesis

$$\mathrm{sign}(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i) \rangle + b) \qquad (5)$$

It can be seen that the kernel and the vector $\boldsymbol{\alpha} = (\alpha_i : i = 1, \ldots, n)$ of Lagrange multipliers define the implementation of Function (5) computed from input space entries $\boldsymbol{x}$ as follows:

$$\mathrm{sign}(\langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle + b) = \mathrm{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right) \qquad (6)$$

According to (4), the aim of this function is to maximize the margin (between classes) and to minimize the training loss. In fact, the sum of the so-called slack variables, $\sum_{i=1}^{n} \xi_i$, is an upper bound of misclassifications of (6) on the training set. It is acknowledged that the Function (6) so achieved has good classification accuracy on unseen cases.

In order to compute the probabilistic outputs, we get rid of the sign function, and we only consider the continuous outputs

$$f_{ac}(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle + b = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \qquad (7)$$

Platt's method then fits a sigmoid to estimate probabilities:

$$h_{ac}(\boldsymbol{x}) = Pr(y = +1|\boldsymbol{x}) = \frac{1}{1 + e^{A_{ac} \cdot f_{ac}(\boldsymbol{x}) + B_{ac}}} \qquad (8)$$

Figure 1 depicts the fit of this sigmoid to the dataset of all patients (2501) at all the available hospitals. Notice that the f_{ac} values follow a bell-shape distribution with most individuals having positive values, which means that they have a survival prediction.

2.3 Optimizing the AUC first

When classification predictions are made comparing the values returned from patients' descriptions $\boldsymbol{x}$ by a rating function with a threshold, as in classification SVM (see Equation (5)), then the performance of these predictions can be assessed using the AUC. According to its probabilistic interpretation, the complementary of this amount (1-AUC) can be used as a loss function. Thus, if g is a hypothesis, its loss evaluated on a test set S' is

$$\begin{aligned} \Delta_{AUC}(g, S') &= Pr(g(\boldsymbol{x}'_i) \le g(\boldsymbol{x}'_j) | y'_i > y'_j) = \\ &= \frac{\sum_{i,j: y'_i > y'_j} 1_{g(\boldsymbol{x}'_i) \le g(\boldsymbol{x}'_j)}}{\sum_{i,j} 1_{y'_i > y'_j}} \end{aligned} \qquad (9)$$

Let us stress that the explicit objective of SVM presented in the preceding section is not to minimize Equation (9). [Cortes and Mohri, 2004] provide a detailed statistical analysis of the difference between maximizing the AUC and minimizing the error rate in binary classification tasks.

Herbrich *et al.* [2000] presented a direct implementation that solves a general ranking problem that is applicable to maximizing the AUC. The core idea is that if a hypothesis $f : \phi(\mathcal{X}) \to \mathbb{R}$ is linear and has to fulfill that $f(\phi(\boldsymbol{x}_i)) > f(\phi(\boldsymbol{x}_j))$, since $y_i > y_j$, then

$$f(\phi(\boldsymbol{x}_i)) > f(\phi(\boldsymbol{x}_j)) \Leftrightarrow f(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)) > 0. \qquad (10)$$

Notice that this statement converts ordering constraints into classification constraints (with one class), but now the input space is $\mathcal{X} \times \mathcal{X}$ and each pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is represented by the difference $\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)$. According to this approach, the aim is to find a hypothesis $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle$ such that $\boldsymbol{w}$ solves the following convex optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}, \xi} \quad & \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i,j: y_i > y_j} \xi_{i,j} \\ \text{s.t.} \quad & \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i) \rangle - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_j) \rangle \ge 1 - \xi_{i,j}, \\ & \xi_{i,j} \ge 0, \quad \forall i, j : y_i > y_j \end{aligned} \qquad (11)$$

For each $\boldsymbol{x}$ of the input space, the hypothesis so found returns

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle = \sum_{y_i > y_j} \alpha_{i,j} (K(\boldsymbol{x}_i, \boldsymbol{x}) - K(\boldsymbol{x}_j, \boldsymbol{x})) \qquad (12)$$

where $\alpha_{i,j}$ are again the Lagrange multipliers computed by the optimizer.

Unfortunately, this approach leads to dealing with one constraint for each element of the dataset

$$\bar{S} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j; +1) : y_i = +1 > y_j = -1\} \qquad (13)$$

whose size is the number of positive (class +1) examples times the number of negatives, #pos×#neg, i.e. $\mathcal{O}(n^2)$ when the size of S is only n. This means that some applications become intractable, although the approach (or a simplified version of it) has been successfully used on other occasions [Joachims, 2002; Bahamonde *et al.*, 2004].

To alleviate the difficulties caused by the size of data sets, it is not straightforward to reformulate Herbrich's approach as an optimization problem with a *small* number of constraints. The main problem is that the loss function (1-AUC) (see Equation (9)) cannot be expressed as a sum of disagreements or errors produced by each input $\boldsymbol{x}_i$.

Following a different procedure, Joachims [2005] recently proposed a multivariate approach to solve this problem with a convex optimization problem that converges using only a few constraints.

The optimization problem is:

$$\begin{aligned} \min_{\boldsymbol{w}, \xi} \quad & \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle + C\xi \\ \text{s.t.} \quad & \langle \boldsymbol{w}, \sum_{y_i > y_j} (1 - y'_{i,j})(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)) \rangle \ge \\ & \qquad \ge \Delta_{AUC}((1, \ldots, 1)(y'_{i,j})) - \xi \\ & \qquad \forall y'_{i,j} \in \{+1, -1\}^{\#pos \cdot \#neg} - \{(1, \ldots, 1)\} \end{aligned} \qquad (14)$$

Despite the enormous potential number of constraints, the algorithm proposed in [Joachims, 2005] converges in polynomial time. Moreover, it only requires a small set of constraints. However, the most interesting result is that the solution $\boldsymbol{w}$ of problem (14) is also the same as that of the optimization problem (11). Additionally, the slack variables in both cases are related by

$$\xi = 2 \sum_{y_i > y_j} \xi_{i,j} \qquad (15)$$

Finally, the multivariate SVM returns a function f_{AUC} of the form

$$f_{AUC}(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle. \qquad (16)$$

Then Platt's method can fit a sigmoid to transform the output of f_{AUC} into a probability.

$$h_{AUC}(\boldsymbol{x}) = Pr(y = +1|\boldsymbol{x}) = \frac{1}{1 + e^{A_{AUC} \cdot f_{AUC}(\boldsymbol{x}) + B_{AUC}}} \qquad (17)$$

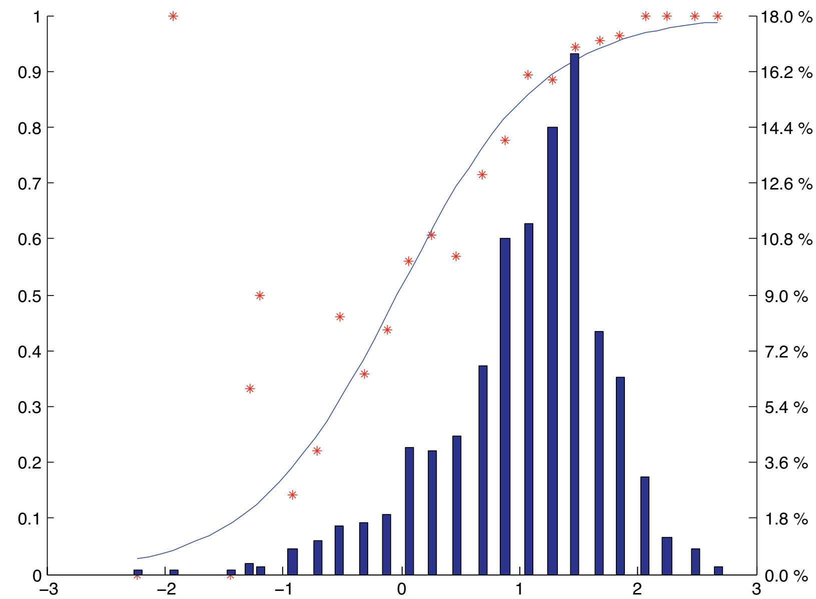

Figure 1: The fit of the sigmoid to the dataset of all patients (2501). The horizontal axis represents the outputs of an SVM. Each '*' mark is the average posterior probability for all examples falling into a bin of width 0.2. The sigmoidal function is the estimation computed by Platt's method [2000] (the output values are labeled on the left vertical side), while the bell-shaped function is the histogram for $Pr(f(\boldsymbol{x}))$ for all the examples (frequencies are labeled on the right).

2.4 Regression is a baseline approach

Considering that probabilities are real numbers, regression algorithms must be a first attempt to learn them. For this purpose, all training examples of class -1 are labeled as 0.

In order to maintain the uniformity of approach with preceding subsections, we considered the regression based on support vectors, therefore we used the so-called Support Vector Regression (SVR). Although there are least squares SVR, we used the standard version; i.e. a learner of a function

$$f_{Re}(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i^- - \alpha_i^+) K(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \qquad (18)$$

where K is once again the rbf (3) kernel, and α_i are the Lagrange multipliers of the solution to the convex optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w},\xi} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_{i=1}^{n} (\xi_i^+ + \xi_i^-), \qquad (19) \\ \text{s.t.} \quad & (\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i) \rangle + b) - y_i \le \epsilon + \xi_i^+, \\ & y_i - (\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i) \rangle + b) \le \epsilon + \xi_i^-, \\ & \xi_i^+, \xi_i^- \ge 0, \qquad i = 1, \ldots, n \end{aligned}$$

However, given that nothing forces f_{Re} (18) outputs to be in $[0, 1]$, we set the hypothesis output to 1 whenever f_{Re} returns values above 1, and 0 for f_{Re} values below 0. In symbols, finally we have the hypothesis

$$h_{Re}(\boldsymbol{x}) = \max\{0, \min\{1, f_{Re}(\boldsymbol{x})\}\} \qquad (20)$$

3 Experimental results

Using a collection of data sets of survival probabilities in critically ill patients, we carried out an experimental comparison of four different learning approaches. SVM followed by Platt's fit of a sigmoid: the accuracy optimizer described in subsection 2.2, which will be represented by SVM(Accu); the multivariate version, aimed at optimizing the AUC (subsection 2.3), for short SVM(AUC); and finally the regression approach, SVR (subsection 2.4). The fourth predictor used was the commercial system APACHE III; we used the customization described in [Rivera-Fernández *et al.*, 1998] that was developed to improve its performance in Spain.

First of all, we have to point out that this is an unfair comparison, since APACHE III was trained with a cohort of 17440 patients from 40 different hospitals in the USA [Knaus *et al.*, 1991]; the Spanish version used records of 10929 patients from 86 ICUs; while the available data sets in our experiments only included 2501 patients. Nevertheless, this comparison is useful to test whether or not the scores achieved by SVM methods are good enough to be considered for future learning tasks.

To estimate the performance of the algorithms described in the preceding section, we used data collected from ICUs at 10 different Spanish hospitals, 6 of which include coronary patients. It is acknowledged among the medical community that coronary diseases generally have a lower mortality risk than other critical illnesses. So from a learning perspective, it makes sense to differentiate between ICUs with and without coronary patients.

The data were organized in 13 different training sets, one

# patients	Hospitals	SVM(AUC) Bs	SVM(AUC) AUC (%)	SVM(Accu.) Bs	SVM(Accu.) AUC (%)	SVR Bs	SVR AUC (%)	*APACHE III* Bs	*APACHE III* AUC (%)
108	1	0.1712	75.82	0.1860	70.60	0.2019	69.86	0.1473	81.76
189	2	0.1887	73.51	0.1998	69.23	0.2444	63.79	0.1710	77.80
194	3	0.1735	75.32	0.1897	65.88	0.1976	70.64	0.1592	78.20
194	4	0.1089	77.20	0.1142	74.93	0.1260	74.35	0.0961	86.17
195	5	0.1102	84.44	0.1094	82.41	0.1078	85.33	0.1079	88.78
239	6	0.1569	74.87	0.1637	69.12	0.1666	71.91	0.1459	77.62
269	7	0.0993	81.09	0.1096	75.75	0.1044	80.47	0.0852	88.02
297	8	0.1205	84.86	0.1277	81.44	0.1301	80.98	0.1127	87.37
337	9	0.1096	81.35	0.1128	77.91	0.1099	79.87	0.1071	81.30
479	10	0.1071	79.32	0.1120	71.74	0.1198	72.74	0.1218	78.22
Averages		0.1346	78.78	0.1425	73.90	0.1509	74.99	0.1254	82.52
919	{2,3,6,8}	0.1494	79.75	0.1500	78.46	0.1546	76.72	0.1432	80.86
1582	{1,4,5,7,9,10}	0.1086	81.79	0.1108	80.37	0.1082	80.08	0.1094	82.63
2501	all	0.1234	81.51	0.1229	81.22	0.1234	80.85	0.1218	82.27

Table 1: Brier scores (Bs) and AUC estimated by a 10-fold cross-validation for the three learners described in the text, and for the commercial system APACHE III. All differences from SVM(AUC) are significant according to a one tail t-test with threshold $p < 0.01$, considering the results on the 10 hospitals. For ease of reading, AUC scores are represented as percentages.

for each single hospital, two collecting the data from not coronary/coronary ICUs respectively, and the last one containing all the data. Each patient in these data sets was described by the same set of variables used by APACHE III. However, given that some of these variables have discrete values, we had to transform them to be handled by SVM-based systems. Thus, we codified each discrete variable using as many new binary variables (with values 0 and 1) as the number of possible values of the original variable, setting only the variable corresponding to the discrete value actually taken by the original variable to '1'.

Performance estimations were made using a 10-fold stratified cross-validation on each of the data sets, for all the algorithms except for APACHE III; since it was already trained with a different data set, we used the available data to test its predictions. Additionally, the data was standardized according to the mean and deviation observed on each training fold.

It is important to recall that the AUC achieved by the Spanish version of APACHE III in our experiments, 82.27% (in percentage) is similar to the amount reported by Rivera-Fernández *et al.* [1998]: 81.82%. This fact supports the representativeness of the sample of critically ill patients considered in the experiments described here.

As usual, when dealing with SVM, the parameter setting stage is very important. To set the regularization parameter C (see optimization problems in Section 2) and the rbf kernel parameter σ (see Eq. (3)) in the three support vector based algorithms, we performed a grid search on a validation set formed by the patients at 3 hospitals: one hospital without coronary patients (8), and 2 with coronary patients (1 and 9); see Table 1. The ranges searched were the following: for C we tested values from 10^{-4} to 10^{2} varying the exponent in steps of 1; for σ we tested values from 10^{-2} to 10^{1} varying the exponent in steps of 0.5. We found that the most promising values were $C = 10^{1}$ and $\sigma = 10^{-2}$ for SVM(Accu) and SVR; and $C = 10^{-1}$ and $\sigma = 10^{-2}$ for multivariate SVM(AUC). It is worth noting that for SVM and SVR the parameter search was aimed at minimizing the Brier score, while for multivariate SVM it was aimed at maximizing the AUC.

Table 1 shows the results obtained (Brier score and AUC) in the experimental setting described above. Focusing on the results obtained by the three support vector algorithms, we can observe that, in general, the best performance (lowest Bs and highest AUC) is achieved by multivariate SVM(AUC). The differences are statistically significant according to a one tail t-test with threshold $p < 0.01$. This should not be surprising for the AUC measure, since this algorithm was specially devised to optimize such a measure. But it also outperforms SVM and SVR in terms of the Brier score, whose parameters were set to optimize this score.

Let us stress that, although the optimization problem posed to SVR is precisely the minimization of the distance between true and predicted probabilities, a large amount of data is required to tie the scores of SVM(AUC) in the Brier score. The underlying reason explaining this behavior may be that the hypothesis space used by SVR is not adequate so as to induce probability distributions from a reduced set of training data, even with an rbf kernel.

As regards the data sets used in the experiments, support vector machines yielded the worst performance on the first three data sets, i.e. the smallest. SVR performance was particularly poor on these data sets. Considering that the rows of Table 1 are in ascending order of size of the data set, the trend indicates that performance could be improved if more training cases were available. In fact, when the data set included all available patients' records, the results obtained were similar to those yielded by APACHE III (recall that it was trained with data sets that were several times bigger). On the other hand, we also observe that survival predictions seem to be slightly harder for ICUs without coronary patients (Hospitals 2, 3, 6 and 8) than for ICUs including coronary patients.

4 Conclusions

We have presented a learning method for estimating probabilities in a real world problem: the prediction of survival in critically ill patients. However, the approach is general enough to be applied to other learning tasks. The method is an alternative to the standard procedure when the learning machine is based on Support Vectors and uses Platt's method [Platt, 2000] to fit a sigmoid. Instead of using an SVM devised to optimize classification accuracy, we propose to use a learner that optimizes the Area Under the ROC Curve (AUC). This can be done using a multivariate SVM described in [Joachims, 2005].

We experimentally compared the results obtained by this method with other approaches, and with a commercial scoring system trained with thousands of cases, APACHE III [Knaus *et al.*, 1991; Rivera-Fernández *et al.*, 1998]. In the reported experiments, we used real data from 10 ICUs at hospitals in Spain that contain records from 2501 patients. The medical description of each patient includes monitoring variables, clinical analysis, and demographic and diagnostic features.

The method proposed here outperforms the standard SVM approach, especially when the available data is scarce, which is the usual situation. On the other hand, increasing the number of training examples reduces differences in performance; even between probability predictions of APACHE III and those made by the baseline method, a simple regression with the output trimmed to the interval $[0, 1]$.

References

[Bahamonde *et al.*, 2004] Antonio Bahamonde, Gustavo F. Bayón, Jorge Díez, José Ramón Quevedo, Oscar Luaces, Juan José del Coz, Jaime Alonso, and Félix Goyache. Feature subset selection for learning preferences: A case study. In Russ Greiner and Dale Schuurmans, editors, *Proceedings of the International Conference on Machine Learning (ICML '04)*, pages 49–56, Banff, Alberta (Canada), July 2004.

[Brier, 1950] G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev*, 78:1–3, 1950.

[Cortes and Mohri, 2004] Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[Hanley and McNeil, 1982] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[Herbrich *et al.*, 2000] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A.J. Smola, P.L. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, Cambridge, MA, 2000.

[Ikeda *et al.*, 2002] Mitsuru Ikeda, Takeo Ishigaki, and Kazunobu Yamauchi. Relationship between brier score and area under the binormal roc curve. *Computer Methods and Programs in Biomedicine*, 67(3):187–194, 2002.

[Joachims, 2002] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[Joachims, 2005] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the ICML '05*, 2005.

[Knaus *et al.*, 1991] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, and A. Damiano. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100:1619–1636, 1991.

[Le Gall *et al.*, 1984] J.R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for ICU patients. *Crit Care Med.*, 12:975–977, 1984.

[Lemeshow *et al.*, 1993] S. Lemeshow, D. Teres, J. Klar, J. S. Avrunin, S. H. Gehlbach, and J. Rapoport. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *Journal of the American Medical Association*, 270(20):2478–2486, November 1993.

[Melville *et al.*, 2005] Prem Melville, Stewart M. Yang, Maytal Saar-Tsechansky, and Raymond J. Mooney. Active learning for probability estimation using jensen-shannon divergence. In *Proceedings of The 16th European Conference on Machine Learning (ECML '05)*, pages 268–279, Porto, Portugal, October 2005.

[Platt, 2000] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A.J. Smola, P.L. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[Provonost and Angus, 2001] P. Provonost and D.C. Angus. Economics of end-life-care in the intensive care unit. *Critical Care Med*, 29(Suppl):46–51, 2001.

[Rivera-Fernández *et al.*, 1998] R. Rivera-Fernández, G. Vázquez-Mata, M. Bravo, E. Aguayo-Hoyos, J. Zimmerman, D. Wagner, and W. Knaus. The APACHE III prognostic system: customized mortality predictions for Spanish ICU patients. *Intensive Care Medicine*, 24(6):574–581, June 1998.