Predicting Mortality of ICU Patients Using Statistics of Physiological Variables and Support Vector Machines

Antonio Bosnjak, Guillermo Montilla

Centro de Procesamiento de Imágenes, Universidad de Carabobo,
Valencia, Venezuela

Abstract

We began using the same variables as SAPS-1 score, adding the rest of variables one by one as recommended by physicians, to observe whether the SVM classification improves. These variables include: Age, HR, SysABP, NISysABP, Temp, RespRate, MechVent, Urine, BUN, HCT, WBC, Glucose, K, Na, HCO3, GCS, and other variables that were added for phase 1: DiasABP, NIDiasABP, Cholesterol, Creatinine, and SaO2. We found a 6.1% error in the Set-A files due to the absence of measures such as: RespRate, Temp, and age. To solve for these errors on phase 1 we chose to input values within the normal range for these physiological variables. We calculated: mean, standard deviation, and range of variation (max and min) for each one of the physiological variables. These values were placed in nodes corresponding to an index and a value of the variable, which were escalated between 0 and 1. We created a matrix where the columns corresponded to: means and standard deviations of the input variables, and rows corresponded to the individual patient's records. We decided to use SVM. Five SVM machines were tested and scored. To conclude, we demonstrate the applicability of SVM for predicting mortality of ICU patients with a final score using set-B of 0.350352 for event 1.

1. Introduction

Risk adjustment systems have been used for more than 20 years in order to predict mortality of Intensive Care Units (ICU); The researchers have developed various methods for scoring: "Acute Physiology and Chronic Health Evaluation" (APACHE IV) [1], Simplified Acute Physiology Score (SAPS III) [2], and Mortality Probability Model (MPM II), [3]. For example, we describe the APACHE IV score developed to improve the accuracy of the APACHE I method, which aims to predict mortality in critically ill patients and to evaluate changes in accuracy relative to APACHE I. For this study, they considered a total of 131,610 patients admitted to ICU during years 2002 and 2003 of which 110,558 met the inclusion criteria. They evaluated 104 intensive care units belonging to 45 hospitals where APACHE III was installed. Zimmerman et al. [1] recalculated the predictions using APACHE IV and obtained better discrimination and calibration, which ought to be used for reference at the ICU's of U.S. We noted that a simplified version of the computerized APACHE system is used in the ICU's of Venezuela.

Lack of proper calibration was observed in subgroups of patients, in most cases. It was often found an underestimation of mortality in low-risk patients and an overestimation of mortality in high-risk patients. It is awkward to explain how a patient with a critical state of physiological variables has survived against his prediction of mortality. This subject is object of continues research, in view to propose further new equations to solve for this mortality model.

The Physionet organization launched the challenge 2012 [4] to the world community to address the problem of Predicting Mortality of ICU Patients: thus, The Physionet/Computing in Cardiology Challenge 2012. The results are now evaluated by an external and impartial team. It tests each program sent to the website over a series of unknown patients. For this challenge 12,000 patients were pooled as three sets of 4,000 patients, called set-A, set-B, and set-C. Each researcher received the set-A with the corresponding results, the Outcomes-a.txt file. Each researcher or group of researchers developed a software program to calculate two results: 1) two classes as: 0 = survivor, 1 = died in-hospital, and 2) The probability of risk of death in-hospital.

We have developed a new method based on Support Vector Machines for calculating Predicting Mortality of ICU, described next.

2. Method

2.1. Data preprocessing

To adjust the data of the 42 variables that are meant to be interpreted by the Support Vector Machine, we

decided to use the first order statistics as the mean and the standard deviation as input vectors for SVM. The variables initially chosen were the same as those used by the SAPS-I Software. This preprocessing method resolves most laboratory variables even if it is lacking on two issues: 1) The physiological variables such as HR, RespRate, Temp, Urine, and Glasgow index are time series that frequently and regularly does vary during patient's stay in the hospital. 2) Most patients' records lack physiological variables or laboratory tests.

Table 1. Ranges of physiological variables.

Variable	Min	Max	Input $\bar{x}$	Units
SAPS-I	2	32		
Age	16	90		years
HR	0	135.54	70	bpm
SysABP	0	181.72	105	mmHg
NISysABP	0	211	105	mmHg
Temp	30.08	38.84	37	ºC
RespRate	0	35.09	10	bpm
MechVent	0	1	0	Boolean
Urine	0	760	0	mL
BUN	3	143	20	mg/dL
HCT	17.63	50.6	32.5	%
WBC	0.1	137.23	10.3	cells/nL
Glucose	47	404.28	120	mg/dL
K	2.98	6.51	4.5	mEq/L
Na	111.5	164.53	138	mEq/L
HCO3	9.77	47	26	mmol/L
GCS	3	15	15	3-15
NIDiasABP	0	97.40	60	mmHg
Cholesterol	84	330	150	mg/dL
Creatinine	0.2	12.95	0.95	mg/dL
SaO2	74	100	89	%
Albumin	0	4.6	0	g/dL

The first problem was solved by calculating the Fourier descriptors of the time series: HR, GCS, RespRate, Temp, and Urine using the following equations. Let $u(n)=x(n)+jy(n)$. We can calculate the Fourier descriptors using the following equation:

$$a(k)=\frac{1}{N}\sum_{n=0}^{N-1}u(n)\exp\left(\frac{-j2\pi kn}{N}\right) \qquad 0\le k\le N-1 \qquad (1)$$

Where the real part and the imaginary part is given by the following equations:

$$\mathrm{Re}[a(k)]=\frac{1}{N}\sum_{n=0}^{N-1}\left\{x(n)\cos\left(\frac{2\pi}{N}kn\right)+y(n)\sin\left(\frac{2\pi}{N}kn\right)\right\} \quad (2)$$

$$\mathrm{Im}[a(k)]=\frac{1}{N}\sum_{n=0}^{N-1}\left\{-x(n)\sin\left(\frac{2\pi}{N}kn\right)+y(n)\cos\left(\frac{2\pi}{N}kn\right)\right\}$$

The first Fourier descriptor matches with the mean value of the signal. The second problem was solved by introducing the normal values for the mean and zero value for the standard deviation. These values were calculated as minimum, maximum, mean, and standard deviation of all physiological variables used. Table I shows the ranges of variation of variables after the evaluation of 4,000 patients of set-A. The third column represents the input value.

2.2. Support vector machines

The SVM by Vapnik [5], is the appropriate learning machine, that minimized the classification error while best finding the hyper-plane of maximum margin that separated the two classes in the featured space. These two classes corresponded as: 0 = survivor or 1=died, in-hospital. The probability risk is calculated using the Gaussian distribution that derives from the distance of classifier to the margin that separates the two classes.

Given a set of points in the input space $\{\mathbf{x_i}\}\subset\Re^n$ $i=1,...,l$ and a function $\mathbf{\Psi}:\mathbf{x}_i\to y_i \quad y_i\in\{-1,1\}$ which assigns to the points one of two possible values. Vapnik [5] proposed projecting the problem to another space (feature space) using a transformation $\Phi:\Re^{\mathbf{n}}\to\Re^m$. In the featured space these classes are linearly separable by a hyperplane of maximum margin. This proposal is presented in figure 1, and the optimization problem is defined by the following equations.

$$\min_{w,b\xi}\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{l}\xi_i \qquad (4)$$

$$y_i\left(\mathbf{w^T\Phi}(\mathbf{x}_i)+b\right)\ge 1-\xi_i \qquad (5)$$

$$\xi_i\ge 0 \qquad i=1,...,l \qquad (6)$$

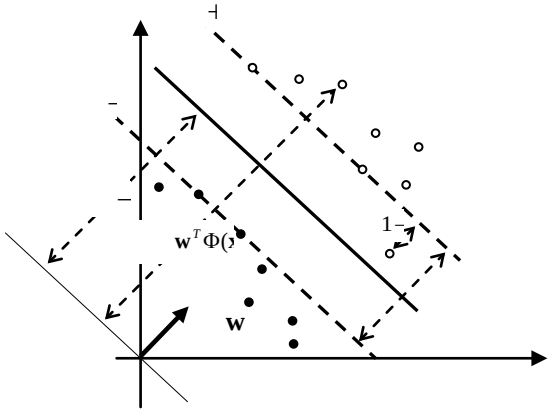

Figure 1. Points and hyperplane in the feature space.

Figure 1 is considered a planar function (distance function) in the featured space. This function is extending in the feature space and it takes zero value over the hyperplane of maximum separation. It can assign a value of +1 to distance function over the nearest points to the optimum hyperplane, which we call border vectors.

Vectors can also be allowed a distance function $1-\xi_i$, which we called outliers. The remaining vectors behind the two planes of distance 1 are called interior points. The variable ***w*** (gradient of the distance function) adjusts to the smoothness of the function. A minimum value of ***w*** gives the maximum smoothness, and a maximum separation between the two classes, since the real distance between the two planes of distance function 1 and -1 is $2/\|w\|$. Equation (5) expresses all the points which are projected behind the planes of distance 1 except the border vectors and the outliers. Equation (4) presents a multi-objective minimization problem that involves the magnitude of ***w*** (coefficient of smoothness or gradient) and the sum of the errors.

Equations (7) - (9) provide the dual problem from the Lagrangian. Equation (7) shows the term $K(\vec{x}_i, \vec{y}_j)$, that represents the scalar product in the featured space. Equation (10) represents the distance function in the feature space, but it can also be plotted in the input space. This is the decision function of the classifier. The zero level surface of this function will be used to solve the modeling problem.

$$\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j \gamma_i \gamma_j K(x_i, y_j) - \sum_{i=1}^{l} \alpha_i \qquad (7)$$

$$0 \le \alpha_i \le C \qquad i = 1,...l \qquad (8)$$

$$\sum_{i=1}^{l} \alpha_i \gamma_i = 0 \qquad (9)$$

$$D(x) = \sum_{i=1}^{l} \alpha_i \gamma_i K(x_i, x) + b \qquad (10)$$

2.3. Training scheme of SVM

In this project, we used the LIBSVM developed on C++ language by Chih-Chung Chang and Chih-Jen Lin [6]. The version 3.12 of this Software is available on their website [6]. For training, we read 4000 files from the set-A. 1000 of these files were chosen for training and 3000 of them were chosen for testing. Figure 2 shows the scheme used by Support Vector Machines (SVM).. This scheme includes the following steps: 1) To read the problem and its pre-processing as explained in the previous section. 2) To scale the vectors between 0 and 1, in order to standardize the physiological variables. 3) The final parameters were selected using a manual process of trial and error. The final parameters were: $\nu = 0.180$, $\gamma = 2.0$, $\varepsilon = 0.58$. 4) We calculated the SVM model and the file of characteristics of physiological variables; these files were distributed with the software package for testing set-B. 5) The test was performed in two ways: a) using the remaining 3000 patients, and b) sending the software through the Website where PhysioNet society procures test making, on the set-B. 6) The test set was scaled. 7) It obtained the prediction results for these tests files using the model obtained during the training phase.

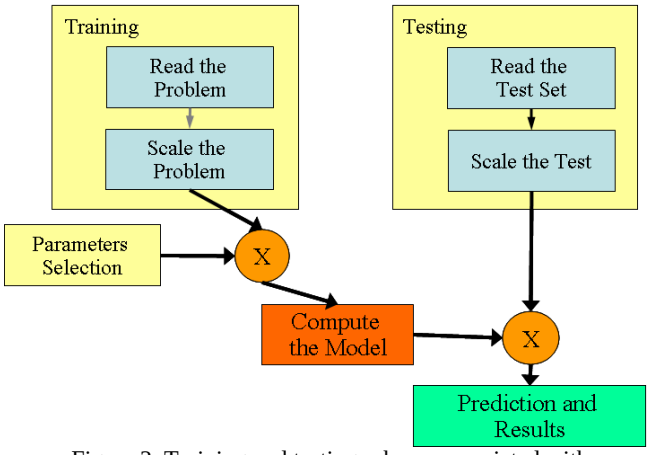

Figure 2. Training and testing scheme associated with Support Vector Machine.

3. Results

The results were evaluated using two score tests: Event 1 defined as the minimum between the Sensitivity (Se) and positive predictivity (+P) and event 2 based on the Hosmer-Lemeshow statistic. Unofficial results for set-A and official results for set-B are:

Table 2. Phase I Results of Challenge 2012.

Set-A	**Entry 1**	**Entry 2**	**Entry 3**
Event 1	0.709386	0.779783	0.815884
Event 2	991.398	1078.99	1055.77
Set-B	**Entry 1**	**Entry 2**	**Entry 3**
Event 1	0.260563	0.278169	0.304577
Event 2	545.662	530.001	659.469

Table 3. Phase II Results of Challenge 2012.

Set-A	Entry 1	Entry 2	Entry 3	Entry 4
Event 1	0.796029	0.853791	0.487365	0.530686
Event 2	38.912	38.912	38.912	38.180
Set-B	Entry 1	Entry 2	Entry 3	Entry 4
Event 1	0.274648	0.297535	0.332746	0.350352
Event 2	35.147	35.147	35.147	35.147

Figure 3 shows the user interface of our program. This interface was created with the objective that the physician can select the most important variables for detecting the mortality risk in the ICU. This interface also allows the engineer the selection: 1) the best parameters of support vector machine, 2) the random selection of a training set, 3) the training, in order to obtain the model, and 4) the test with the remaining set-A patients.

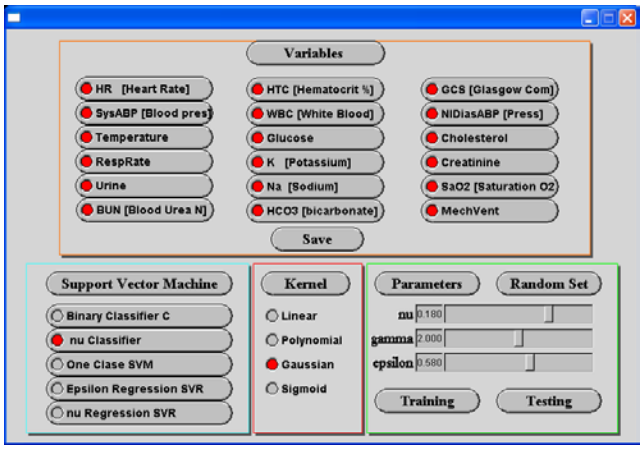

Figure 3. User interface to select the variables and parameters of the best SVM machine. This is designed by authors.

Figure 4 depicts the graph corresponding to the calculation of Hosmer - Lemeshow statistics. It was evaluated for each patient on set-A where we obtained a score of 0.530686 for event 1 and 38.180 for event 2. On this graph the number of deaths observed of ICU match with the number of deaths predicted by our software. A Small change can be seen at the curves around 0.2 when the prediction probability becomes a higher.

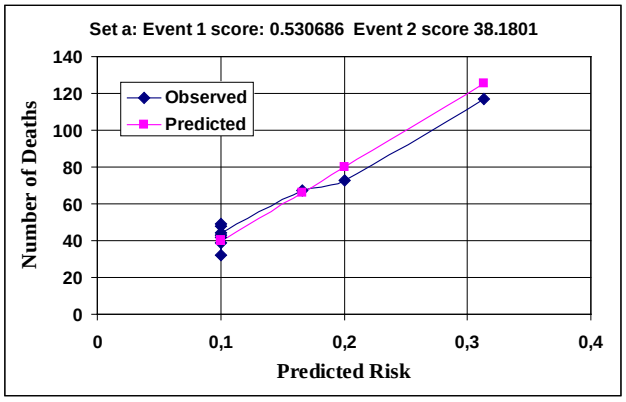

Figure 4. Calcule of Hosmer – Lemeshow statistics for set-A of the final Entry.

4. Conclusions

We observed an over-fitting of SVM to set-A because the score to event 1 is 0.8158 while the same evaluation using set-B is 0.3045. For phase 2 we are set to improve the training strategy of SVM; separating set-A by a random pattern, in order to correct over-fitting of SVM, and including different variables and scores.

For Phase II, 1000 patients were selected for training and 3000 unknown patients for testing set-A. The system was improved, and we reduce the over-fitting of SVM machine. Our results show that the final score for event 1 is 0.530686. And the final score with set-B testing by PhysioNet society is 0.350352. It demonstrated that the jump between the set-A of training set and testing with set-B was significantly reduced. Finally our last software "entry 4" of Phase II was tested with the set-C obtaining a score of 0.3333. This shows we needed a continues research in order to improve the prediction of mortality of ICU.

References

[1] Zimmerman J E, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine 2006;* 34:1297-1310.

[2] Metnitz PGH, Moreno RP, Almeida E, Jordan B, Bauer P, Abizanda CR, Iapichino G, Edbrooke D, Capuzzo M, Le-Gall JR. SAPS 3 – From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive Care Medicine 2005;* 31:1336-1344.

[3] Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III)*. *Critical Care Medicine 2007;* 35: 827-835.

[4] National Institutes of Health NIH. Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012. [Accessed: August 27, 2012]. Available from: http://physionet.org/challenge/2012/

[5] Vapnik V. Statistical Learning Theory. Wiley. 1998.

[6] Chang Ch, Lin CJ, LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. [Accessed: September 1, 2012]. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Address for correspondence.

Name: Antonio BOSNJAK SEMINARIO
Address: Centro de Procesamiento de Imágenes. Universidad de Carabobo. Final Av. Universidad, Bárbula.
Ciudad: Valencia. Edo. Carabobo. Venezuela
E-mail: antoniobosnjak@yahoo.fr