A Neuro-Fuzzy Based Alarm System for Septic Shock Patients with a Comparison to Medical Scores

Jürgen Paetz[1,2], Björn Arlt[1,2]

[1] J.W. Goethe-Universität Frankfurt am Main,
Fachbereich Biologie und Informatik,
Institut für Informatik, AG Adaptive Systemarchitektur
Robert-Mayer-Straße 11-15, D-60054 Frankfurt am Main, Germany
[2] Klinikum der J.W. Goethe-Universität,
Klinik für Allgemein- und Gefäßchirurgie
Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany
`{paetz,arlt}@cs.uni-frankfurt.de`
`http://www.medan.de`

Abstract During the last years we collected data of abdominal *septic shock* patients from clinics all over Germany. The mortality of septic shock is about 50%. Septic shock is related to immune system reactions and unusual measurements. Septic shock patients are intensely medicated during their stay at the intensive care unit. To help physicians recognizing the critical states of their patients as early as possible, we built a *rule based alarm system* based on a neuro-fuzzy inference machine. Analysing the patient data in a time window, we show the time dependency of the classification results. We give detailed classification results and explanation by rules. The results are compared to results obtained by using the most common *scores* in intensive care medicine. We discuss the advantages of the paradigms "neural networks" and "scores", and we answer the important question: Is a neural network more performant than scores for abdominal septic shock patient data?

1 Introduction

Septic shock is of prime importance in intensive care medicine. Epidemiologic investigations of septic shock patients show the high risk potential and the extensive therapy situation in intensive care units (ICU) [1]. Variables are often investigated as isolated variables, not as a multidimensional whole, e.g. a recent study inspects the role of thrombocytes [2].

Our approach to reduce mortality of septic shock patients is the automated, intelligent retrospective search of information in documented patient records. We analysed the data of 138 patients by using most of the usually documented metric variables (e.g. blood pressure, leukocytes, medicament doses). Data was collected in German hospitals from 1998 to 2001. Up to now we have digitized 138 collected handwritten patient records. 70 of the 138 patients are deceased

(50.7%). Our analysis of metric data carries on the analyses already done with another data base and other methods, e.g. [3]. The scores that are often used in the ICU are described in Sect. 2. To find interesting rules within the high number of all the rules coming from subsets of all the variables we used a neuro-fuzzy algorithm based on [4, 5] which is described in Sect. 3. Subsequently, in Sect. 4 achieved results are presented. We compare the scoring results to the neuro-fuzzy results, and we give some meaningful rule examples. The results culminate in an alarm system whose performance is analysed.

2 Scores in Intensive Care Medicine

How could physicians assess the patient's health status as objectively as possible? An easy practiced method is to model expert opinions by using a *score*, i.e. a sum of points. This action is not fully objective, but it represents joint expert opinions. By defining a threshold the score can be utilized as a classifier for outcome prediction. We present shortly the most common scores used in intensive care medicine. Classification results of applied scores are given in Sect. 4.4.

a) *SOFA* (Sepsis-Related Organ Failure Assessment) [7]: the SOFA score assesses organ malfunctions by whole-numbered values. The sum of these values for the single organs is called SOFA score.

b) *APACHE II* (Acute Physiological and Chronic Health Evaluation) [8]: APACHE II is a score for outcome prognosis of ICU patients with respect to acute disorders, age and the overall health status (0 to 71 whole-numbered points).

c) *SAPS II* (Simplified Acute Physiology Score) [9]: The SAPS II score is a variable reduced APACHE II score. Only 13 instead of 34 variables are used.

d) *MODS* (Multiple Organ Dysfunction Score) [10]: The MODS score assesses organ states (lungs, liver, kidney, haemogram, heart, neurological system) by whole-numbered points.

In the SOFA and MODS score we do not include the Glasgow Coma Scale (GCS) for assessing the neurological state because of its impreciseness and its high subjectivity.

3 The Neuro-Fuzzy System

The supervised neuro-fuzzy algorithm [4] uses the class information of the data within its adaptation process. Here, we use the outcome labels "survived" and "deceased" for the classes. The main advantages of the algorithm are:

- The training uses a simple heuristic geometric adaptation process that softens the combinatorical explosion (exponential growth) during the rule generation process due to multiple dimensions.
- Irrelevant attributes for every rule are detected. This is the case if a part of a rule R has the format "**if** ... **and** var$_j$ **in** $(-\infty, +\infty)$ **and** ... **then** class ...". Then, the value of variable j is not relevant and so the variable could be omitted, leading to a shorter rule R.

– Adaptive learning of the exact shape of the trapezoid membership functions.

Let us describe the ideas of the algorithm. The 2-layer neural network has neurons in the hidden layer with n-dimensional asymmetrical trapezoidal fuzzy activation functions. Every neuron in the first layer belongs to only one class and represents a fuzzy rule. During the learning phase these neurons p are adapted, i.e. the sides of the upper, smaller rectangles (= *core rules*) and the sides of the lower, larger rectangles (= *support rules*) of the trapezoids are adapted to the data. For every new training data point x of class c this happens in four phases, initialized by the first training sample x_1 for which one neuron is committed with infinite side expansions in every dimension for the support rule and no (zero) expansion for the core rule (core rule = x_1):

1. *cover*: if x lies in the region of a support rule of the same class c as x, expand one side of the corresponding core rule to cover x and increment the weight of the neuron,
2. *commit*: if no such support rule covers x, insert a new neuron p at point x of the same class and set its weight to one and its center $z := x$. The expansions of the sides of the support rule – associated with the new neuron – are set to infinite; the expansions of the sides of the core rule – associated with the new neuron – are set to zero,
3. *shrink committed neuron*: for a committed neuron shrink the volume of the support (and the core rectangle if necessary) within one heuristically chosen dimension of the neuron in relation to the neurons belonging to other classes,
4. *shrink conflict neurons*: for all the neurons belonging to another class $\neq c$, heuristically shrink the volume of both rectangles of these neurons within one dimension in relation to x.

A sketch of the standard algorithm [4] is given in Appendix A. For implementation details of our more technical modifications and improvements see [5, 6]. Here, we place emphasis on the application to our septic shock patient data.

4 Results

At first we give a short description of the database and the datasets. Then, we present the experimental conditions and the classification results of the neuro-fuzzy and score classifiers with a discussion. Meaningful rule examples are presented.

4.1 The Datasets

Our database consists of 138 septic shock patients. The metric data that we consider is composed of daily measurements and doses of medicaments. For the experiments that are presented in Fig. 1, we consider different *periods of time*: F3 (first 3 days of ICU stay), S3 (first 3 days after the septic shock occurence), ALL (all days of ICU stay), D6–8 (days 6,7 and 8 counted from the last day, i.e. day 0, of ICU stay), D2–4 (days 2,3 and 4 counted from the last day of ICU stay), L5, L3, L2, L1 (last 5 resp. 3, 2, 1 day(s) of ICU stay). In Figs. 2 and 3 we regard only the period L3, cf. the explanation in Sect. 4.3.

We consider the following datasets in our contribution: frequent16 (the most frequent 16 measured variables), haemogram, heart, lungs, bac (breathing and catecholamines), bpt (systolic and diastolic blood pressure, thrombocytes) and the single variables systolic blood pressure, diastolic blood pressure, thrombocytes, see appendix B for more details.

Here, we will mention only our main preprocessing steps [3]: sampling (mean values of 24h) and missing value removal (missing values are replaced by random values from a normal distribution within the interval of the so called interquartile range, abbr. IQR, with the median as the center).

4.2 Experimental Conditions

All the samples of the datasets are classified by the neuro-fuzzy system. Training was done with 50% of the samples and testing with the remaining 50%. No data from training patients is used for testing (disjunct patient sets). All experiments with one dataset were repeated five times, so that all given results are average values of all experiments. Thus, random results are avoided.

To compare classification performance, we use the area under the ROC curve (AUC). The ROC curve is given by sensitivity values on y-axis and specifity values on x-axis. It is equal to 0.5 if the classifier has no performance (random classification) and equal to 1.0 if the classifier performs without errors. Here, each ROC curve is calculated using 10 different classification thresholds, i.e. sensitivity/specifity settings. The AUC is calculated with the well known trapeze rule for numerical integration.

4.3 Neuro-Fuzzy-System Performance

The best outcome predictor would be one that warns the physician at the first day of ICU admission or at the first day of septic shock appearance (that is mostly the second day of the patient's ICU stay). With the dataset "frequent16" we show in Fig. 1 that it is not possible to train an adequate prediction system within such an early time interval (F3 AUC = 0.53, S3 AUC = 0.58). Also it is not reliable to base the system on the samples of *all* days (ALL AUC = 0.67). Unfortunately, within the last two days often less samples are measured by physicians so that the results considering L2 and L1 are not trusting or significant. The best classification results are achieved considering L3 (AUC = 0.92). Thus, in the following we consider only the period L3.

The results achieved with the dataset "frequent16" (period L3) are encouraging, but it is not reliable for physicians to key in 16 variables. The SOFA score is based only on 10 variables. Is it possible to achieve a similar performance

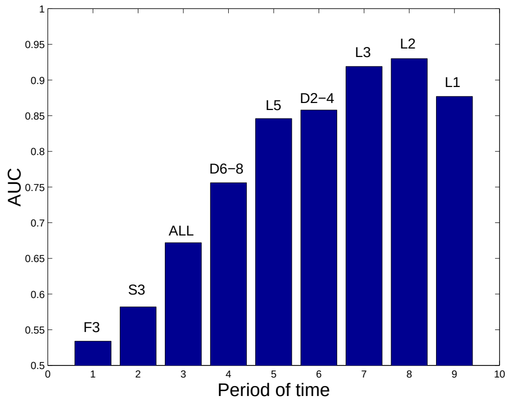

Figure 1. "Frequent16" data: Area under ROC curve (AUC) for different periods of time of the ICU stay.

using less variables? To answer this question, we tried out different compositions of variables. The results are presented in Fig. 2. We see that single variables have not a sufficient performance. The best performance is achieved by the haemogram (AUC = 0.90) and by "frequent16" (AUC = 0.92). As said before keying in 14 resp. 16 variables in an alarm system is too much. Thus, the system "bpt" (3 variables) with an equal AUC = 0.90 is a good candidate for building an alarm system, but only if it performs not worse than the established scores, cf. the next section.

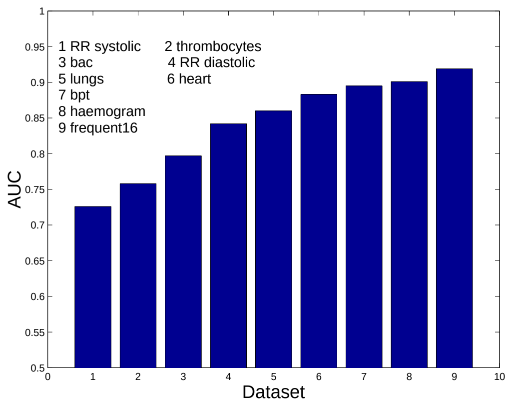

Figure 2. Area under ROC curve (AUC) for different data sets (last three days of ICU stay).

4.4 Score Performance

In Fig. 3 we see that the three scores MODS, SAPS II and APACHE II perform almost equally well (AUC = 0.79 , 0.79 resp. 0.80), considering the time period L3. We achieve an obvious better classification using the SOFA score (AUC = 0.89). The SOFA score, that is composed of 10 variables, does not perform better than the "bpt" or "frequent16" system. Therefore, it could be replaced in the ICU by using an alarm system whose alarm behaviour is described in Sect. 4.6.

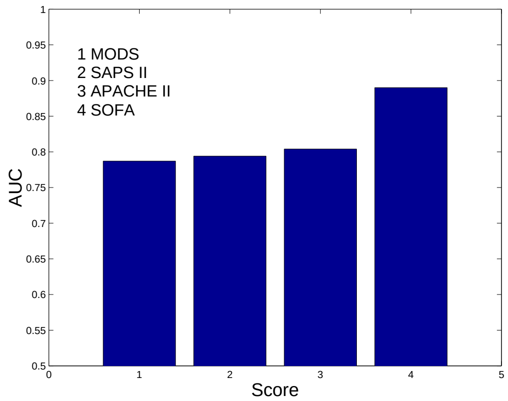

Figure 3. Area under ROC curve (AUC) for the scores MODS, SAPS II, APACHE II and SOFA.

One question remains open: Why is the adaptive data driven neural network approach not *much* more performant than a "simple" score? Firstly, only one score – the SOFA score – has a similar performance. Secondly, the SOFA score was evaluated on 1643 patients in the USA (more than 10 times the number of patients that we consider), surely including many expert opinions or even statistical methods. Thirdly, our analyses are retrospective analyses with a lower quality as potential prospective analyses due to missing values.

Without presenting all the details of our analysis, the results of a binary logistic regression, using SPSS 9.0, are very similar to the results achieved by our rule-based system. Thus, binary logistic regression is also not much more performant than the SOFA score on our data.

Finally, the SOFA score is computed somewhat similar to a neural network output: 1 up to 4 points are given individually for every variable state by a nonlinear, discrete step function. Then, these values are added up. In a neural network the activities of the neurons in the first layer are added up in the second layer. Of course, the neural network learns its activities by a data driven training autonomously. This would give better results as a score if and only if the step functions within a score are chosen badly. Thus, it seems that the step functions in the SOFA score are chosen well. If our data would have more complex, nonlinear class borders, a neural network approach would surely give much better results since then one cannot guess correct classification without a machine learning approach, even not by investing a lot of time for trial and error based data analysis. Inventing a score without using a machine learning approach is experience based trial and error data analysis, hoping to obtain good results.

To sum up, we cannot beat sigificantly the SOFA score's performance by our approach (or by binary logistic regression without rule generation), but we are able to reduce the number of variables and give important insight by our rules. We need three variables only for our alarm system instead of ten variables used for calculation of the SOFA score. It seems that the variables in the "bpt" system are the "core" variables whose trends are common to most of the patients. The use of additional variables in the other systems seems to be more influenced by individual patient behaviour with less performant classification results. Besides the "bpt" system, a mixture of variables from different systems as in the "frequent16" system or in the SOFA score give good results.

4.5 Generated Rules

We present the results of our rule generation for the dataset "bpt". On average we generated 7.6 rules for the class "deceased" and 8.2 for the class "survived". To evaluate the performance of the rules we calculated the frequency (percentage of all samples that imply the rule) and the confidence (percentage of samples of the correct class considering all samples that imply the rule) of the rules on the test data. Let us give two important examples of frequent and confident (support) rules:

1) class **survived** with frequency = 30.5% and confidence = 98.4% if
systolic blood pressure $\geq$ 111.8 and
diastolic blood pressure $\geq$ 41.7 and
thrombocytes in (264.0,700.0)

2) class **deceased** with frequency = 40.5% and confidence = 91.4% if
systolic blood pressure $\leq$ 127.5 and
diastolic blood pressure $\leq$ 62.8 and
thrombocytes $\leq$ 282.0

These rules show that a lower systolic and diastolic blood pressure and a lower number of thrombocytes indicate very critical diseasedness. Quantified results as the results above may lead to more precise therapy options in future.

4.6 Resulting Alarm System

We identified the "bpt" system as a performant classifier (considering the time period L3). In fact, our aim is not to predict the patients' outcome within the last three days of their ICU stay. We want to warn the physician *during* the *entire* ICU stay if a patient is critical. To set up the alarm system we proceed as follows: We tried out different output thresholds for our neuro-fuzzy system, i.e. thresholds for adjusting sensitivity/specifity considering all samples from the time series of the patients. For every sample we generate the alarm "very critical" – using the trained neuro-fuzzy system – if the (normed) output o_d for the class "deceased" is $\geq \kappa_1$ resp. the alarm "critical" if the (normed) output o_d for the class "deceased" is $\kappa_1 > o_d \geq \kappa_2$ with chosen thresholds $\kappa_1 > \kappa_2$. If o_d is $< \kappa_2$ no alarm is given. In this manner we use our outcome predictor, trained within the period L3, as an alarm system for the entire ICU stay. Every time a patient becomes as critical as most of the deceased patients in the last three days, the system generates an alarm.

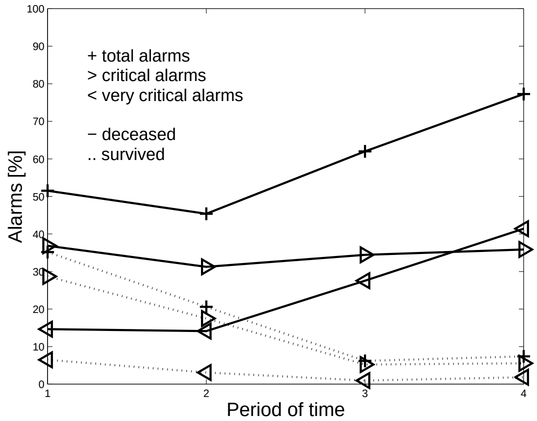

Figure 4. Progression of alarms over time. Critical, very critical and total alarms for deceased and survived patients. Mean values over test patients. Periods: 1 = first three days (F3), 2 = first half of ICU stay, 3 = second half of ICU stay, 4 = last three days (L3).

In Fig. 4 we present the average percentage of alarms given for the patients for different time periods. Ten times more alarms in the last three days for deceased patients than for survived patients seems to be reliable for a bedside system.

5 Conclusion

Our aim was the extraction of information from septic shock patient measurement data. For this purpose we applied an efficient improved neuro-fuzzy algorithm to generate rules. We obtained interesting rules for the classes of deceased and survived patients. A detailed comparison of the classification performance of scores showed that the best score for septic shock outcome diagnosis is the SOFA score. We identified the systolic and diastolic blood pressure/thrombocytes ("bpt") system as the most relevant for outcome prediction.

It is possible to achieve a similar classification performance as by the SOFA score, but with less variables (3 instead of 10) together with a rule-based explanation. Our alarm system produces reliable alarms (in the last three days of the ICU stay ten times more alarms for deceased patients then for survivors). Finally, in April 2002 we started a multicenter study to check the clinical usefulness of our system.

Acknowledgement: Our work is supported by the German Research Foundation (DFG). The authors thank all the participants of the MEDAN working group.

References

1. Hanisch, E., Encke, A.: Intensive Care Management in Abdominal Surgical Patients with Septic Complications, published in: E. Faist (ed.) Immunological Screening and Immunotherapy in Critically Ill Patients with Abdominal Infections. Springer-Verlag (2001) 71–138
2. Vanderschueren, S. et al.: Thrombocytopenia and Prognosis in Intensive Care. Crit. Care Med. **28** (2000) 1871–1876
3. Paetz, J., Hamker, F., Thöne, S.: About the Analysis of Septic Shock Patient Data. 1st Int. Symp. on Medical Data Analysis. LNCS Vol. 1933. Springer-Verlag (2000) 130–137
4. Huber, K.-P., Berthold, M.R.: Building Precise Classifiers With Automatic Rule Extraction. Proc. of the IEEE Int. Conf. on Neural Networks **3** (1995) 1263–1268
5. Paetz, J.: Metric Rule Generation with Septic Shock Patient Data. Proc. of the 1st IEEE Int. Conf. on Data Mining (2001) 637–638
6. Brause, R., Hamker, F., Paetz, J.: Septic Shock Diagnosis by Neural Networks and Rule Based Systems. In: M. Schmitt et al. (eds.) Computational Intelligence Processing in Medical Diagnosis. Physica-Verlag (2002) 323–356
7. Vincent, J.-L. et al.: The SOFA (Sepsis-Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure. Intensive Care Med. **22** (1996) 707–710
8. Knaus, W.A. et al.: APACHE II: A Severity of Disease Classification System. Crit. Care Med. **13**(10) (1985) 818–829
9. Le Gall, J.R. et al.: A New Simplified Acute Physiology Score (SAPS II) Based on a European / North American Multicenter Study. The Journ. of the Am. Med. Assoc. **270** (1993) 2957–2963
10. Marshall, J.C. et al.: Multiple Organ Dysfunction Score: A Reliable Descriptor of a Complex Clinical Outcome. Crit. Care Med. **23**(10) (1995) 1638–1652

Appendix A – The Neuro-Fuzzy Algorithm

parameters: p_i^s (i-th neuron of class s), w_i^s (weight of p_i^s)
1. reset weights:
for s = 1 **to** number of classes **do**
  for i = 1 **to** m_s (number of neurons for class s) **do**
    $w_i^s := 0$;
    set core rule volume of p_i^s to zero;
  end
end

2. consider all samples $(x; c)$ with c as class label of x:
for all samples x **do**
  if p_i^c covers x **then**

3. cover:
  $w_i^c := w_i^c + 1$;
  adjust core region, so that it covers x;

4. commit new neuron:
  else
  $z_{m_c+1}^c := x$;
  set core rule volume of $p_{m_c+1}^c$ to zero;
  set support rule volume to infinity;

5. **for** $s \neq c$, $1 \leq j \leq m_s$ **do**
    shrink $p_{m_c+1}^c$ using z_j^s;
  end
  $m_c := m_c + 1$;
  $w_{m_c}^c := 1$;
  end

6. shrink conflict neurons:
  for $s \neq c$, $1 \leq j \leq m_s$ **do**
    shrink p_j^s using x;
  end
end

Appendix B – The Datasets

We use the following abbreviations: CVP = central venous pressure, PTT = partial thromboplastin time, TPT = thromboplastin time, AT = anti thrombin, EK = erythrocytes concentrate, FFP = fresh frozen plasma, I:E = inspiratory:expiratory (pressure). The units in the following datasets are only mentioned once:

frequent16: heart frequency [1/min], systolic blood pressure [mmHg], diastolic blood pressure [mmHg], temperature [°C], CVP [mmHg], O_2 saturation [%], leukocytes [1000/μl], haemoglobin [g/dl], haematocrit [%], thrombocytes [1000/μl], PTT [s], sodium [mmol/l], potassium [mmol/l], creatinin [mg/dl], blood sugar [mg/dl], urine volume [ml].

haemogram: leukocytes, erythrocytes [1000/μl], thrombocytes, TPT [%], PTT [s], haemoglobin, haematocrit, thrombin time [s], AT3 [%], fibrinogen [mg/dl], total protein [g/dl], blood sugar [mg/dl], EK [ml], FFP [ml].

heart: heart frequency, systolic blood pressure, diastolic blood pressure, CVP, cristalloids [ml], colloids [ml], adrenaline [μg/kg/min], noradrenaline [μg/kg/min], dopamine [μg/kg/min], dobutamine [μg/kg/min].

lungs: arterial pO_2 [mmHg], arterial pCO_2 [mmHg], base excess [-], bicarbonat [mmol], O_2 saturation, O_2 medication [l/min], Peak [cmH_2O], I:E [-], respiratory frequency [1/min], FiO_2 [%], PEEP [mmHg].

bac: FiO_2, PEAK, respiratory frequency, adrenaline, noradrenaline, dopamine, dobutamine.