Journal of Intelligent and Fuzzy Systems, Special Issue, Vol. 12(1): Challenges for Future Intelligent Systems in Biomedicine, IOS Press 2002, pp.69-78

Rule Generation and Model Selection Used for Medical Diagnosis

Jürgen Paetz, Rüdiger Brause
J.W. Goethe-Universität Frankfurt am Main,
Fachbereich Biologie und Informatik, Institut für Informatik,
Robert-Mayer-Straße 11-15, D-60054 Frankfurt am Main, Germany

Abstract In medical data analysis classification combined with rule generation is a common technique to obtain diagnosis results together with a rule based explanation. In this contribution we apply a neural network based rule generator in the domain of septic shock research. The septic shock is of special interest in intensive care medicine due to its high lethality rate. We describe the functionality of the neuro-fuzzy algorithm and present classification and rule generation results of our analysis. Because we repeated our analysis with randomly selected test data to calculate statistically valid mean results, we generated one neural network with different architecture for each repetition. To decide the important question which of the different models should be used in the application phase, we propose a useful method based on similarity measures for rules resp. rule sets to select one representative network out of the set of trained networks.

1 Introduction

The use of artificial neural networks has become a powerful, widely-used technique for classification analysis of medical data, see [1, 2, 3, 4].

Standard neural network techniques like backpropagation do not explain their classification results by *rules*. Particularly physicians are interested in such rules to get insight in the classification process, e.g. to draw conclusions for therapy. Thus, scientists have developed methods that allow the generation of rules within the classification process [5, 6, 7, 8, 9] or the extraction of rules after the classification process [10].

The septic shock is one of the most common reasons of death in intensive care units (ICUs). One of our goals is the application of a knowledge based method to the analysis of septic shock patient data. For this paper, we chose the algorithm [5, 6] in its improved version [11], see Sect. 2. Our analyses are restricted to abdominal intensive care patients who developed a septic shock during their stay at the ICU. The abdominal septic shock has a high lethality rate in the ICU up to 50%. Some more details are described in Sect. 3.

We repeat all the experiments with randomized partitions of the medical data into training and test data to get meaningful, statistically valid results. We evaluate our classification results with standard performance measures, e.g. classification error on training and test data sets. The particular rules are evaluated with a frequency and confidence measure (Sect. 3).

Because we repeated our experiments generating more than one different neural networks resp. rule bases we developed a new index to choose one of the architectures based on similarity measures. The index will enable us to identify the network that is most similar to all

the others, the so called *representative* network, see Sect. 4. This is a new and very important method that should be used after the training procedure: Naturally, by repeating the training with randomized partitions of the data into training and test data you will obtain different network models with a varying classification and rule performance. Less data quality yields in even higher performance variations.

Which model has the best diagnostic characteristics *and* the best explanatory rule set? The performance of each network depends on the data partition, so it is not possible to choose a "best" network, e.g. the network with the highest classification performance on the special test data since the high classification performance may be merely a random effect. Moreover the model with the highest classification performance may be composed of too many rules, or it may be composed of not very well interpretable rules. The other way round, the network built by the smallest number of rules need not to have a sufficient classification performance. Nevertheless, for a real application – like our medical application – we need to choose one particular model. With our new index we support data analysts choosing one representative, particular model out of a larger set of models, considering the rule performance and rule structure of the different models. Finally, in Sect. 5 we discuss our results.

2 Metric Rule Generation

Our main goals are the generation of rules for septic shock patient data (Sect. 3) and the model selection (Sect. 4). For the convenience of the reader we shortly describe the ideas of the algorithm [5, 6] and discuss its (dis-)advantages. Since the improvement and implementation details [11] are not relevant for the model selection in Sect. 4, we will only present a sketch of the original algorithm without discussing the improvements. The addressed details could be found in [11, 12].

2.1 The Neuro-Fuzzy Algorithm

The supervised neuro-fuzzy algorithm [5, 6] uses the class information of the data within its adaptation process. Here, we use the outcome labels $\{$survived, deceased$\}$ for the classes. Principal 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.
- 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 resulting in a shorter rule R.
- Adaptive learning without stating membership functions a-priori.
- No rule aggregation of generated rules after rule learning is required.
- Starting the training with a-priori known rules after fuzzification is possible.
- Extraction of both crisp and fuzzy rules is possible, see Sect. 2.2.

Let us describe the ideas of the algorithm. The 2-layer network in Fig. 1 has neurons in the hidden layer with n-dimensional asymmetrical trapezoidal fuzzy activation functions (see Fig. 2). Every neuron in the first layer belongs to only one class and represents a fuzzy rule.

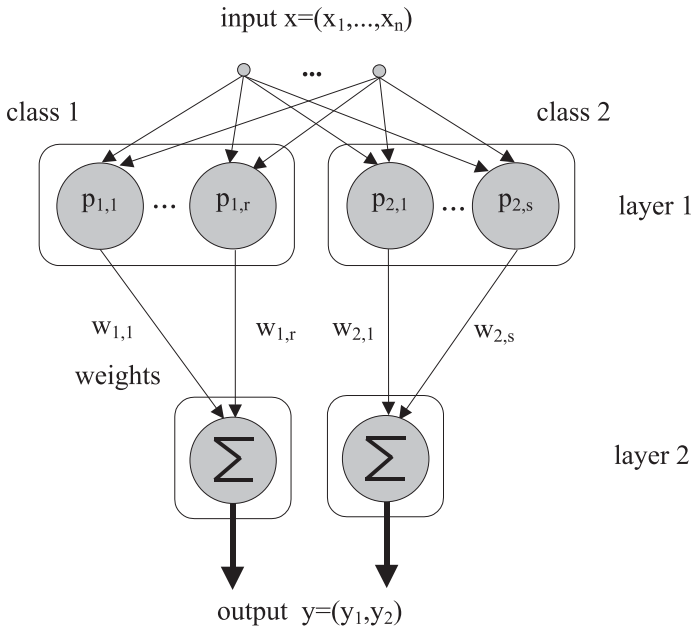

Figure 1: The neural network structure of the rule generation algorithm for two classes. In the first layer the network contains neurons separately for every class, i.e. the first layer is not fully connected to the second layer. The $w_{1,i}$'s denote the r weights for class 1 neurons; the $w_{2,i}$'s denote the s weights for class 2 neurons.

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 for which one neuron is committed with infinite side expansions in every dimension:

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 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 within one dimension in relation to x.

At the beginning of each entire training cycle all the weights are set to zero. Classification is done by a winner-takes-all mechanism, i.e. calculate the activity $s_i(x, c_i)$ as the sum of the weights multiplied by fuzzy activation for every class c_i and choose the class $c_{\max}$ as classification result, where $c_{\max} := \mathrm{class}(\max_{c_i}(s_i(x, c_i)))$.

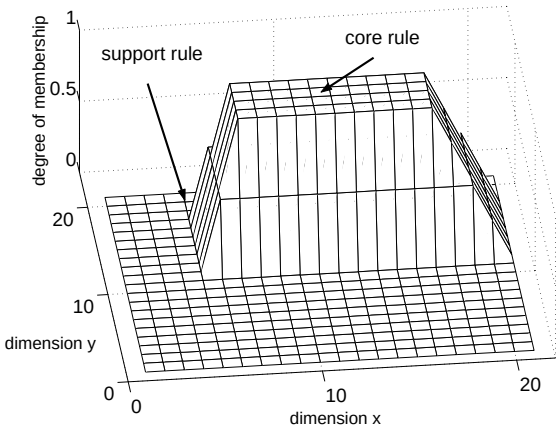

Figure 2: A 2-dimensional asymmetric trapezoidal membership function, interpreted as a core rule (small rectangle at the top of the trapezoid) and a support rule (larger rectangle at the bottom of the trapezoid) in the algorithm.

Negative aspects of the algorithm that weaken the quality of the rules, especially in high dimensional problems, are:

- the relation to presentation order of the training samples with an unfavourable expansion of core rules,
- the immediate creation of new rules for outlier data,
- the large overlapping of support rules,
- the extensive overlapping of core rules with different class labels that may cause semantic confusion of the rules coming from different classes.

In principle the overlapping of rule regions is reasonable and desired to achieve fuzzy rules. But the rules in algorithm [5, 6] tend to overlap too intensely due to the heuristic cover-commit-shrink-procedure. We addressed to these problems by some modifications of the cover-commit-shrink procedure [11] which are not discussed here. As already mentioned, in this contribution we place emphasis on *model selection* that could be used both for the neuro-fuzzy model [5, 6] and [11].

2.2 Some Basics About Rules

Now, let us define the common rule performance measures *frequency* and *confidence* similar to [13] where the measures are used for evaluating association rules. These measurements are used in the context of neuro fuzzy systems in [11, 12]. These measures will play a major role in our rule generation process. We define them in definition 1. Before, we give the definition of a *rectangular rule* extracted from a fuzzy rule. In our case it will be better to use crisp, rectangular rules than fuzzy rules due to their higher precision, see [14]. In medical applications fuzzy rules might be better understandable if the problem is a small and easy problem. However, here we have to deal with more precise information for the more complex problem “septic shock”.

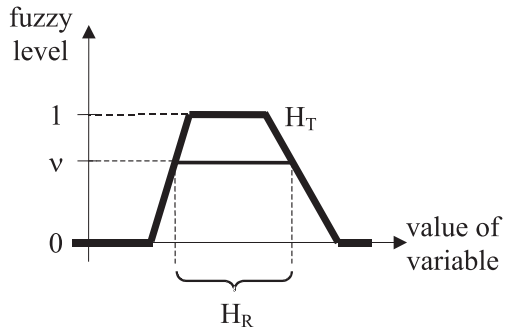

Figure 3: A 1-dimensional membership function, interpreted as a trapezoid H_T and a 1-dimensional hyper-rectangle (= interval) H_R, cut from H_T at fuzzy level ν.

Definition 1: (Rectangular Rule)
Let F be a fuzzy rule generated by the algorithm and H_T the corresponding hyper-trapezoid. A **rectangular rule** *R is defined by a hyper-rectangle H_R that is cut from H_T at a chosen fuzzy level (degree of membership), see Fig. 3. H_R is allowed to have infinite side expansions in some dimensions.*

Definition 2: (Frequency and Confidence)
Let R be a rectangular rule and H_R the corresponding hyper-rectangle, cf. definition 1. The rule R is associated to one class k.

a) The **class** s **frequency** $\text{freq}_s(R)$ *of the rule R of class k is defined as the number of samples of class s that lie in H_R divided by the number of all the samples in the whole dataset. If $s = k$ we say shortly* **frequency** $\text{freq}(R)$ *instead of class k frequency.*

b) The **class** s **confidence** $\text{conf}_s(R)$ *of the rule R of class k is defined as the number of samples of class s that lie in H_R divided by the number of all the samples that lie in H_R. If $s = k$ we say shortly* **confidence** $\text{conf}(R)$ *instead of class k confidence.*[1]

Only rules R that are sufficient *frequent* and *confident*, those with $\text{freq}(R) \geq \min_{\text{freq}}$ and $\text{conf}(R) \geq \min_{\text{conf}}$ using a-priori defined thresholds $\min_{\text{freq}}$ and $\min_{\text{conf}}$, are interesting for a presentation to physicians. These thresholds must be high enough to provide interesting, significant rules and low enough to generate a sufficient number of rules. Thus, an expert of the application area – in medical applications a physician – has to be involved in order to design proper thresholds for useful results. Then, sufficient frequent and confident rules may be a great benefit for physicians. Of course, different thresholds could be defined for different classes to warrant more flexibility.

Usually, rectangular rules cut at a fuzzy level $\nu_1 > \nu_2$ are more confident and less frequent. If the rectangular rules cut at a high fuzzy level get too small they may loose their statistical significance and so they may get less confident [11].

[1] Multiplied by 100 the measures can be interpreted as a percentage.

3 Application to Septic Shock Patient Data

In abdominal intensive care medicine patients are in a very critical condition. Often patients develop a *septic shock*, a phenomenon that is related to mechanisms of the immune system [15, 16, 17, 18] and which is still an important research subject for medical experts and data analysts because there are no ultimately satisfying results published until now. The septic shock is associated with a high lethality of about 50%. It is always related to measurements leaving the normal range (e.g. blood pressure, temperature, respiratory frequency, number of leukocytes), and it is often related to multiorgan failure. The epidemiology of 656 intensive care unit patients is elaborated in an older study made from 1995 to 1997 at the Klinikum der J.W. Goethe-Universität Frankfurt am Main [19].

In Sect. 3.1 we describe the data and in Sect. 3.2 we present our results of the rule generation and classification.

3.1 The Data

The multicenter data that we used was collected in several German clinics from 1997 to 2000. The data base D consisted of 138 patients in August 2001 and is still increasing. 50.7% (70 patients) of the septic shock patients deceased.

We will at first report results of two different experiments with the dataset F_{16} composed of the 16 most frequent measured variables, i.e. the variables: heart frequency [1/min], systolic blood pressure [mmHg], diastolic blood pressure [mmHg], temperature [$^\circ$C], central venous pressure (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]. We analyzed the data

- of the three days after the septic shock appears for the first time (dataset F_{16}^{first}) and
- of the last three days of the patient's stay at the ICU (dataset F_{16}^{last}).

Preparing the data for analysis, we used some preprocessing steps similar to [20, 21]: All values from patient records were sampled in 24h intervals due to their different, unregular measure frequency. If there were different values for one variable available within a 24h interval then the mean of those values was used for analysis. Due to this preprocessing, some short-time dynamics might be lost but the data values got very stable for our analysis. We required that 13 of the 16 variables were measured for each sample. Thus, for F_{16}^{first} there remained 394 out of 411 and for F_{16}^{last} there remained 303 out of 413 samples. The remaining samples were allowed to have a maximum of 3 missing values that were replaced by a random value from the interquartil range (IQR) of the distribution of the variable for which the value was missing. All the experiments were done with a randomized partition of the data into 50% training and 50% test data, such that the test data set contained no samples of patients of the training data set and vice versa.

3.2 Results

In Table 1 we see that the results are obviously much better on the last three days than on the first three days, i.e. a better classification performance with a less number of rules. Additional experimental results have shown us that the classification performance (ROC area) increases

Table 1: Correct classifications on training and test data with standard deviation. ROC area (0.50 means worst, 1.00 means best classification performance). Number of rules. Mean results of 5 repetitions.

	training correct [%]	std. [%]	test correct [%]	std. [%]	ROC area	rules
F_{16}^{first}	65.89	8.28	56.83	1.52	0.58	34.2
F_{16}^{last}	88.68	2.27	84.02	4.44	0.92	16.2

approximately linearly from the first day to the last day of the patient's stay at the ICU. Thus, an early warning system has a small diagnostic capability if trained only with data of the first ICU day. It is common sense that phycians know already on the last day that the patient is likely to survive or decease. Therefore, it is a compromise to build an alarm system with data e.g. from the last five days, so that there is a benefit of an alarm for the physician and a scope for treating the patient.

We give two examples for rules that we have generated, one for the class deceased and one for the class survived, coming from the dataset F_{16}^{last}, cut at fuzzy level 0.0 (= support rules):

I) "**if** diastolic blood pressure ≥ 59.52 **and** thrombocytes ≥ 27.00 **and** potassium ≤ 4.70 **then** class survived **with** test data frequency 0.40 **and** test data confidence 0.84 **from** 33 different test data patients"

II) "**if** systolic blood pressure ≤ 139.14 **and** CVP ≥ 4.61 **and** haematocrit ≤ 34.70 **and** thrombocytes ≤ 271.50 **and** creatinin ≥ 0.71 **then** class deceased **with** test data frequency 0.29 **and** test data confidence 0.93 **from** 24 different test data patients"

In our situation a transformation of the rules into a fuzzy notation with linguistic variables like "high", "normal", "low" is not very helpful for a physician – although possible – because these linguistic variables have no predefined medical sense. A lot of ideas concerning the issue "crisp or fuzzy rules" are discussed in [14].

4 Model Selection

Repeating the rule generation process five times, we have generated five different network models for the experiments in Sect. 3. So the question is: which model should we use in an application phase? Our aim is not to choose one model out of different kinds of models, e.g. decision trees, neural networks, Bayes networks etc. Once you have chosen the model paradigm that is the most suitable for your data in practice you still have the problem to choose one particular model from the models generated by the same paradigma by different training and test data partitions. No systematic approach is known to the authors where this important problem is addressed. A fundamentally different approach is the *combination* of all five networks, for example by ensemble averaging [22] or boosting [23], that we will not address here.

It makes no sense to choose the model with the best classification and/or rule performance. It may be the case that such a model has good classification and/or rule performance only with

the special test and training data partition that was used to build and test the model. It is also not useful to choose the model with the lowest number of rules for the same reason. All these criteria are arbitrary, misleading and influenced by random effects. The model with the highest classification performance (usually calculated on test data) may have a too high number of rules with a lower explanatory power. The model that is composed of a small number of rules with a high explanatory power may have less classification performance.

Our approach considers all the important factors together for choosing a model, i.e. the structure of the rule sets of the different models, the rule performance and the number of rules. Then our aim is finding the model that is most similar to all the other models to use it as a system prototype. This model will be the favorized model, the *representative* model. In the next Sect. 4.1 we will define an index based on similarity measures for this task. We discuss a preliminary benchmark experiment on the IRIS dataset in Sect. 4.2. In Sect. 4.3 we present our results for our medical data.

4.1 Similarity Measure - Representative Rule Set

We introduce stepwise the *rule similarity measure* for two rules, the *rule set similarity measure* for two rule sets and the *similarity placement index (SPI)*. With the SPI we choose the *representative* rule set.

Rules R – coming from the algorithm in Sect. 2 – have in general the form: "**if** var_1 **in** (a_1,b_1) **and** ... **and** var_n **in** (a_n,b_n) **then** class c", i.e. intervals as antecedents and a class label as conclusion. For technical reasons n-dimensional data was transformed to the hypercube $[0,1]^n$. Then, $d_i^{(l)}$ and $d_i^{(r)}$ in definition 3b) are well defined. In general a_i could be $-\infty$ and b_j could be ∞. If $a_i = -\infty$ and $b_i = \infty$ then the variable var_i is not relevant for the rule and could be omitted.

Definition 3: (Similarity)
a) Let R_1 and R_2 be two rules. Define rel_{R_i} as the number of relevant variables of rule R_i, $i = 1, 2$. Define rel_{id} as the number of variables, that are relevant for R_1 as well as for R_2. Then, we define a first simple similarity measure for two rules as:

$$sim_1(R_1, R_2) := \frac{rel_{id}}{\max\{rel_{R_1}, rel_{R_2}\}} \quad . \tag{1}$$

$sim_1 \in [0,1]$ compares the number of relevant variables of two rules.

b) Formula (1) does not take into account the different extensions, i.e. interval borders of the rules. Let i be the index of a relevant variable, that is relevant for R_1 as well as for R_2. Let (a_i, b_i) resp. (c_i, d_i) the associated intervals of the support rules with regard to the i-th relevant variable for R_1 resp. R_2. For $i = 1, \ldots, rel_{id}$ we define:

$$d_i^{(lr)}(R_1, R_2) := d_i^{(l)} + d_i^{(r)} \quad , \tag{2}$$

where $d_i^{(l)}$ resp. $d_i^{(r)}$ denote the differences of the left resp. right interval borders, defined

by:

$$d_i^{(l)} := \begin{cases} 0.5 & , a_i = c_i = -\infty \\ 0.5 \cdot (1 - |c_i - a_i|) & , a_i \neq -\infty, c_i \neq -\infty \\ 0 & , (a_i \neq -\infty \wedge c_i = -\infty) \vee (a_i = -\infty \wedge c_i \neq -\infty) \end{cases} \tag{3}$$

and

$$d_i^{(r)} := \begin{cases} 0.5 & , b_i = d_i = \infty \\ 0.5 \cdot (1 - |d_i - b_i|) & , b_i \neq \infty, d_i \neq \infty \\ 0 & , (b_i \neq \infty \wedge d_i = \infty) \vee (b_i = \infty \wedge d_i \neq \infty) \end{cases} . \tag{4}$$

The value $d_i^{(l)} = 0.5$ resp. $d_i^{(r)} = 0.5$ stands for no difference; $d_i^{(l)} = 0$ resp. $d_i^{(r)} = 0$ stands for the maximum difference. We combine these distances with sim_1 by a weighted sum and normalize it in order to define an extended similarity between two rules:

$$sim_2(R_1, R_2) := \frac{sim_1(R_1, R_2) + \frac{\varrho}{rel_{id}} \sum_{i=1}^{rel_{id}} d_i^{(lr)}(R_1, R_2)}{1 + \varrho} . \tag{5}$$

By parameter $\varrho \in [0, 1]$ the influence of the interval border differences to sim_2 could be varied.

c) Let R_1, R_2 be two rules of a rule set R. In the last step we finally define $sim_3 : R \times R \longrightarrow [0, 1]$ as the **rule similarity measure**, considering additional interesting **influence functions** $G_1, \ldots, G_g$, $g \in \mathbb{N}$ with image $\text{im}(G_i) = [0, 1]$ for the application. In our example later on we will set G_1 as the confidence and G_2 as the frequency of two rules R_1 and R_2.

$$sim_3(R_1, R_2) := \frac{sim_2(R_1, R_2) + \sum_{j=1}^{g} \kappa_j (1 - |G_j(R_1) - G_j(R_2)|)}{1 + \sum_{j=1}^{g} \kappa_j} . \tag{6}$$

With the help of the weighting parameters $\kappa_j \in [0, 1]$, $j = 1, \ldots, g$, the influence of the influence functions could be varied.

d) Now we define a rule set similarity measure for two rule sets. Let $R = \{R_1, \ldots, R_m\}$ and $\hat{R} = \{\hat{R}_1, \ldots, \hat{R}_n\}$ be two rule sets and $S = (s_{jk}) := (sim_3(R_j, \hat{R}_k))$ the matrix of all rule similarity measures $sim_3(R_j, \hat{R}_k)$ of rules from rule set R and $\hat{R}$. Let c be the number of classes and $\sharp R_c$ the number of rules per class in rule set R. Then, we define the **rule set similarity measure** of two rule sets $Sim(R, \hat{R}) \in [0, 1]$ using the highest rule similarity measure for two rules from different classes and the number of rules in the rule set:

$$Sim(R, \hat{R}) := \frac{\frac{\sum_{j=1}^{m} max_k\{s_{jk}\} + \sum_{k=1}^{n} max_j\{s_{jk}\}}{m+n} + \sum_{l=1}^{c} \tau_l \cdot \frac{min\{\sharp R_l, \sharp \hat{R}_l\}}{max\{\sharp R_l, \sharp \hat{R}_l\}}}{1 + \sum_{l=1}^{c} \tau_l} . \tag{7}$$

With the help of the weighting parameters $\tau_l \in [0, 1]$, $l = 1, \ldots, c$, the influence of the numbers of rules for different classes could be varied.

Table 2: SPI's for the 10 models generated with dataset IRIS and placement for model selection. For each model the correct training and test classification percentage is stated. $\sharp R_c$ indicates the number of rules of class c.

	SPI	placement	training correct	test correct	$\sharp R_1$	$\sharp R_2$	$\sharp R_3$	$\sharp R_1 + \sharp R_2 + \sharp R_3$
model 1	0.6885	7	88.00	94.67	1	3	4	8
model 2	0.5960	9	96.00	97.33	1	1	2	4
model 3	0.6830	8	96.00	98.67	1	3	3	7
model 4	0.4690	10	100.00	92.00	1	1	1	3
model 5	0.7273	2	97.33	94.67	1	3	2	6
model 6	0.7285	1	97.33	94.67	1	3	2	6
model 7	0.7098	5	94.67	97.33	1	3	3	7
model 8	0.7220	3	97.33	93.33	1	3	3	7
model 9	0.7147	4	97.33	93.33	1	2	2	5
model 10	0.7089	6	96.00	90.67	1	3	3	7
mean	0.6748	–	96.00	94.67	1.0	2.5	2.5	6.0

We consider only values s_{jk} in (7), coming from rules of the same class. The rule similarity measure of rules from different classes is set to 0, i.e. the computation of Sim in (7) becomes easier, if the maxima of s_{jk} are searched only within identical classes.

e) Finally, let there be r rule sets $R^{(1)}, \ldots, R^{(r)}$. Then we call a rule set $R^{(j)}$ the **representative rule set**, if the rule set has the maximum value of all SPI_i's:

$$j = index(\max_{i=1,\ldots,r}\{SPI_i\}) \tag{8}$$

with the **similarity placement index (SPI)**

$$SPI_i = SPI(R^{(i)}) := \frac{\sum_{j=1, j\neq i}^{r} Sim(R^{(i)}, R^{(j)})}{r-1} \quad , \tag{9}$$

i.e. the representative rule set is the rule set that is in mean the most similar one to all the other rule sets, the one with the highest SPI.

We believe that by changing the similarity measures or choosing other influence functions the ideas of the SPI are useful also in other data analysis models, not only in our special neuro-fuzzy system. The parameters ρ, κ_j, τ_l have to be chosen in relation to the individual, problem dependent significancy of the influence functions.

In the research area of *case based reasoning* [24] representative samples (samples=cases) are calculated using similarity measures to build sample classifiers. Interestingly, we could interpret our representative rule set as a kind of "best case" if we consider the rule sets as cases, even though the analogy is not complete.

4.2 Benchmark Results for the IRIS Dataset

We show that not only a small rule number but also a high classification performance need not to be the same as a representative rule structure ("high SPI"). In Table 2 the SPI's for the wellknown IRIS dataset [25] is shown. The IRIS data is 4-dimensional (numerical variables). It has 3 classes and contains 150 measurements. As proposed before, we randomly selected

Table 3: SPI's for the 5 models generated with dataset F_{16}^{last} and placement for model selection. For each model the correct training and test classification percentage is stated. $\sharp R_c$ indicates the number of rules of class c.

	SPI	placement	training correct	test correct	$\sharp R_1$	$\sharp R_2$	$\sharp R_1 + \sharp R_2$
model 1	0.5688	3	86.58	90.91	7	9	16
model 2	0.5397	5	86.58	85.07	6	7	13
model 3	0.5745	1	89.10	78.91	10	7	17
model 4	0.5693	2	89.10	82.31	9	7	16
model 5	0.5676	4	92.05	82.90	9	10	19
mean	0.5640	–	88.68	84.02	8.2	8.0	16.2

50% of the data for training and 50% for testing. We believe that the common relevant variables of sim_1 should have the highest influence on the overall similarity. The influence functions and the rule numbers are used for finetuning. Thus, we set the parameters to: $\rho = 0.1$, $\tau_l = 0.1$ for both classes, $\kappa_1 = 0.1$ (frequency of the rules, calculated on the test data), $\kappa_2 = 0.1$ (confidence of the rules, calculated on the test data).

We see that the training and test performance and the number of rules of each model differ from other models' performance. The difference of the SPI for model 5 and 6 is due to the different rule performance values of the confidence and the frequency and different rule borders of the 6 rules. Model 2 and 4 have the smallest number of rules (4 resp. 3 rules). These models have the lowest placement. Model 2 and 4 have both neither the best nor the worst test classification performance. Model 3 has the best test classification performance using the higher number of 7 rules. Its placement is only 8. The reason may be the less representative rule structure of model 3 compared to the other models. The model with the highest SPI is model 6 with 6 rules and 94.67% test classification performance. This is a well-balanced total performance regarding classification and rule number results.

As a final remark we state that the only common feature of all trained models is the sufficiency of only one rule for class 1.

4.3 Results for Septic Shock Patient Dataset

We calculated the rule and rule set similarity measures and the similarity placement index (SPI) on our 5 models that we have generated on the dataset F_{16}^{last}. In our experiments we set the parameters to: $\rho = 0.1$, $\tau_l = 0.1$ for both classes, $\kappa_1 = 0.1$ (frequency of the rules, calculated on the test data), $\kappa_2 = 0.1$ (confidence of the rules, calculated on the test data).

In Table 3 we noted the SPI's for the dataset F_{16}^{last}. Of course, the mere values of the SPI's are not interesting for the reader. But the result is very important: The models with the smallest (model 2) or the highest (model 5) rule number have less test performance than model 1 which is the model with the highest test performance. Here, the model with the lowest test performance – but with a higher training performance – has the highest SPI. In this case the criteria "highest test classification performance", "highest rule number", "lowest rule number" and "highest SPI" would select different models. This is an indication that a well adjusted rule structure ("highest SPI") is different to other criteria. Even with the lowest test performance on a special data division we argue that it is reliable to choose such a network for an application.

Thus, by using the SPI's we can choose a system that is not randomly performant on a special data partition – especially if the results are inhomogeneous – but the system that is most likely to be the most representative one for our problem, the system with a representative rule structure.

5 Conclusion

We have applied the rule generation neuro-fuzzy algorithm [5, 6] in our improved version [11] for rule generation. Beside the main ideas of the algorithm [5, 6] and its (dis-)advantages we described how we could generate crisp, precise, rectangular rules from fuzzy rules. This may be helpful in a lot of applications where too vague fuzzy variables do not provide the necessary precision, see also [14].

As an application of our data analysis tools we have experimentally showed that data of septic shock patients underlie a dynamic process: The data sampled within the first days at the ICU or within the first days of septic shock appearance is not well classifiable, but it become much better classifiable if it is sampled within the last days at the ICU. Some rule examples are given to point out the usefulness of the rectangular rules.

By our approach it will be possible to train a rule based system with data of the last days of ICU stay where the classes are separable. Of course, physicians need a system that warns as early as possible, but this is no contradiction. In the application phase data will be entered and results could be requested on every day of the patient's ICU stay, although it will be more likely that a warning is given from the system at the end of the ICU stay. We will examine the details of a reliable warning system with our data base as soon as more patients are documented.

Two questions could be asked concerning the dataset F_{16}: Do we need 16 variables or are there important variables that contain already the necessary information? Is the classification with a neuro-fuzzy method more successful than a common medical score like SOFA [26]? We work on these questions and expect very interesting results.

One important technical aspect that we have invented and applied is the model selection based on similarity measures: If we have n different models of the same kind, what model should we use in an application? We presented an approach that is based on rule and rule set similarity measures. This measures were adapted especially to our rule generation algorithm but they could easily be adapted to other data analysis algorithms using specific influence functions, e.g. for other kinds of neural networks. We hope that this idea is very useful for all scientists who arbitrarily chose one model in the past. Our new model selection strategy avoids random effects caused by different partitions of training and test data by detecting a representative rule set. It could be worthwhile to search for even more objective similarity measures and parameter settings.

Acknowledgement: The work was done within the DFG-project MEDAN (German Research Foundation, Ref. no. HA 1456/7-2). The authors like to thank all the participants of the MEDAN working group especially Prof. Hanisch and Dr. Holzer for providing our work with medical background knowledge. We thank Dipl.-Inform. Arlt and all the medical doctoral students in the working group for their help on the data base and also for discussing unclear data sets. In the end we thank the reviewers for helpful comments.

References

[1] P.J.G. Lisboa, E.C. Ifeachor, P.S. Szczepaniak, eds., Artificial Neural Networks in Biomedicine, Springer-Verlag, London, 2000.

[2] W. Baxt, Application of Artificial Neural Networks to Clinical Medicine, Lancet 346 (1995), 1135–1138.

[3] W. Penny, D. Frost, Neural Networks in Clinical Medicine, Med. Decis. Making 16 (1996), 386–398.

[4] T. Villmann, Neural Networks Approaches in Medicine - a Review of Actual Developments, in: Proc. of the 9th European Symp. on Artificial Neural Networks (ESANN), Bruges, Belgium (2000), 165–176.

[5] K.-P. Huber, M.R. Berthold, Building Precise Classifiers with Automatic Rule Extraction, in: IEEE Int. Conf. on Neural Networks (ICNN), Perth, Australia 3 (1995), 1263–1268.

[6] M.R. Berthold, K.-P. Huber, From Radial to Rectangular Basis Functions: A New Approach for Rule Learning from Large Datasets, Internal Report 15-95, Univ. Karlsruhe, Germany, 1995.

[7] D. Nauck, R. Kruse, Obtaining Interpretable Fuzzy Classification Rules From Medical Data, Artificial Intelligence in Medicine 16(2) (1999), 149–169.

[8] R. Brause, F. Friedrich, A Neuro-Fuzzy Approach as Medical Diagnostic Interface, in: Proc. of the 9th European Symp. on Artificial Neural Networks (ESANN), Bruges, Belgium (2000), 201–206.

[9] B. Fritzke, Incremental Neuro-Fuzzy Systems, in: Proc. SPIE's Optical Science, Engineering and Instrumentation: Applications of Fuzzy Logic Technology IV, San Diego, USA 3165 (1997), 86–97.

[10] H. Tsukimoto, Extracting Rules From Trained Neural Networks, IEEE Transactions on Neural Networks 11(2) (2000), 377–389.

[11] J. Paetz, Metric Rule Generation with Septic Shock Patient Data, in: Proc. of the 1st IEEE Conf. on Data Mining (ICDM), San Jose, USA (2001), 637–638.

[12] R. Brause, F. Hamker, J. Paetz, Septic Shock Diagnosis by Neural Networks and Rule Based Systems, in: M. Schmitt et al., eds., Computational Intelligence Techniques in Medical Diagnosis and Prognosis, Physica-Verlag, Heidelberg, 2002, pp. 323-356.

[13] R. Agrawal, R. Skrikant, Fast Algorithms for Mining Association Rules, in: Proc. of the 20th Int. Conf. on Very Large Databases (VLDB), Santiago de Chile, Chile (1994), 487–499.

[14] W. Duch, A. Rafal, K. Grabczewski, A New Methodology of Extraction, Optimization and Application of Crisp and Fuzzy Logical Rules, IEEE Transactions on Neural Networks 12(1) (2000), 277–306.

[15] E. Hanisch, A. Encke, Intensive Care Management in Abdominal Surgical Patients with Septic Complications, in: E. Faist, ed., Immunological Screening and Immunotherapy in Critically Ill Patients with Abdominal Infections, Springer-Verlag, Berlin, 2001, pp. 71–138.

[16] A.M. Fein et al., eds., Sepsis and Multiorgan Failure, Lippincott Williams & Wilkins, Baltimore, 1997.

[17] E. Neugebauer, D. Rixen, M. Raum, U. Schäfer, Thirty Years of Anti-Mediator Treatment in Sepsis and Septic Shock – What have We Learned? Arch Surg 383 (1998), 26–34.

[18] R.M. Hardaway, A Review of Septic Shock, Amer Surg 66(1) (2000), 22–29.

[19] S. Wade, M. Büssow, E. Hanisch, Epidemiology of SIRS, Sepsis and Septic Shock in Surgical Intensive Care Patients, Chirurg 69 (1998), 648–655.

[20] S. Tsumoto, Rule Discovery in Large Time-Series Medical Databases, in: Proc. of the 3rd European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), Prague, Czech Republic (1999), 23–31.

[21] J. Paetz, F. Hamker, S. Thöne, About the Analysis of Septic Shock Patient Data, in: Proc. of the 1st Int. Symp. on Medical Data Analysis (ISMDA), Frankfurt am Main, Germany (2000), 130–137.

[22] U. Naftaly, N. Intrator, D. Horn, Optimal Ensemble Averaging of Neural Networks, Network 8(3) (1997), 283–296.

[23] Y. Freund, R.E. Schapire, Experiments with a New Boosting Algorithm, in: Proc. of the 13th Int. Conf. on Machine Learning (ICML), Bari, Italy (1996), 148–156.

[24] J. Kolodner, Case-Based Reasoning, Morgan Kaufmann Publishers, San Mateo, 1993.

[25] R.A. Fisher, The Use of Multiple Measurements in Axonomic Problems, Annals of Eugenics 7 (1936), 179–188, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/.

[26] J.-L. Vincent et al., The SOFA (Sepsis-Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure, Intensive Care Medicine 22 (1996), 707–710.