Fuzzy modeling based on Mixed Fuzzy Clustering for health care applications

Marta C. Ferreira, Cátia M. Salgado, Joaquim L. Viegas, Hanna Schäfer, Carlos S. Azevedo,
Susana M. Vieira and João M. C. Sousa
IDMEC, Instituto Superior Técnico, Universidade de Lisboa
{marta.ferreira, joaquim.viegas, carlos.azevedo, susana.vieira, jmsousa}@tecnico.ulisboa.pt
catiasalgado@live.com, schaeferhannaj@gmail.com

***Abstract*—This papers proposes two novel approaches for the identification of Takagi-Sugeno fuzzy models with time variant and invariant features. The proposed Mixed Fuzzy Clustering algorithm is proposed for determining the parameters of Takagi-Sugeno fuzzy models in two different ways: (1) the antecedent fuzzy sets are determined based on the partition matrix generated by the Mixed Fuzzy Clustering algorithm; (2) the input features are transformed using the same algorithm and the antecedent fuzzy sets are derived using Fuzzy C-Means clustering. The proposed approaches are tested on four different health care applications: readmissions in intensive care units, administration of vasopressors and mortality. The results show that the proposed clustering algorithm resulted in an increase of the performance of the fuzzy models in three out of four applications in comparison to the use of Fuzzy C-Means.**

I. Introduction

Intensive care unit (ICU) data has grown exponentially in the last decades, making the ICU a particularly appealing setting for the implementation of data-based systems [1]. Such systems would acquire large quantity of data to discover hidden associations and understand patterns and trends in medical data for diagnostic, prognostic and therapy [1], [2]. However, due to issues related to the collection and analysis of data, the available data is still underutilized for the care of the critically ill. Fuzzy modeling provides transparent predictive models and linguistic interpretations of the decision making process, showing great potential in dealing with vague information. Hence, it is especially well suited for health care applications since it can provide clinical insight into the classifier structure.

Most medical databases are characterized by different types of data including constant features (numerical, categorical and binary attributes) representing information such as age or gender, and time series, such as the changes in time of physiological measurements like blood pressure. Thus, data-based models should be able to simultaneously consider both types of data. However, there is a lack of alternatives to simultaneously handle constant features and time series data, and traditional methods are not suitable to deal with the complexity of datasets with different data types.

Time series clustering has been shown useful in providing knowledge in various domains. In [3], a thorough overview of time series clustering techniques is presented. Spatiotemporal clustering is a process of grouping objects based on their spatial and temporal similarity [4]. In [5], the Fuzzy C-Means (FCM) clustering technique is augmented for spatiotemporal clustering of data with spatial and temporal features.

In the present work, the spatiotemporal clustering concept is extended to any data set containing both time variant and time invariant features (mixed datasets). The algorithm is applied to health care data to predict the outcome of critically ill patients.

This paper introduces the novel Mixed Fuzzy Clustering (MFC) algorithm in Section II and presents Takagi-Sugeno fuzzy models (FM) in Section III. Details of the methodologies used for modeling based on MFC are provided in Section IV, while the experimental results for each study case are demonstrated in Section V. Finally, main conclusions are presented in Section VI.

II. Mixed Fuzzy Clustering

MFC is a novel clustering method based on Fuzzy C-Means [6] which deals with both time variant and invariant features. This method introduces a generalization of the spatiotemporal concept to any set of time variant and time invariant features and its extension to the analysis of multiple time-series. In this approach each sample x_i, with $i = 1, ..., n$, is characterized by features that are constant during the sampling time in analysis and by features that change over time (multiple time-series).

$$x_i = (\mathbf{x}_i^s, X_i^t) \tag{1}$$

In order to extend the spatiotemporal clustering method proposed in [5] which only deals with one time-series to the case of multiple time-series, a new dimension is introduced, to handle p time variant features. As presented in equations 2 and 3, the static component of the samples is represented by $\mathbf{x}_i^s$, where r is the number of static features, and the temporal component of the samples is represented by the matrix X_i^t with number of columns equal to the number of temporal features p and rows equal to the number of temporal samples q.

The MFC algorithm clusters the dataset using an augmented form of the FCM. The main difference between the augmented and the classical FCM relies on the distance function. In the augmented FCM a new pondering element λ is included, factoring the importance to be given to the time variant component. The distance is also calculated separately for each time-series.

$$\mathbf{x}_i^s = (x_{i,1}^s, ..., x_{i,r}^s) \tag{2}$$

$$X_i^t = \begin{pmatrix} x_{i,1,1}^t & x_{i,1,2}^t & \cdots & x_{i,1,p}^t \\ x_{i,2,1}^t & x_{i,2,2}^t & \cdots & x_{i,2,p}^t \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,q,1}^t & x_{i,q,2}^t & \cdots & x_{i,q,p}^t \end{pmatrix} \quad (3)$$

The static prototypes are represented for each cluster l by v_l^s or $\mathbf{v}_l^s$ and are computed in equation 4, for $l = 1, ..., c$.

$$\mathbf{v}_l^s = \frac{\sum_{i=1}^n u_{l,i}^m \mathbf{x}_i^s}{\sum_{i=1}^n u_{l,i}^m} \quad (4)$$

The temporal prototypes for each cluster l and feature k are represented by $v_{l,k}^t$ or $\mathbf{v}_{l,k}^t$, computed following equation 5 and the matrix of temporal prototypes for cluster l is represented by V_l^t.

$$\mathbf{v}_{l,k}^t = \frac{\sum_{i=1}^n u_{l,i}^m \mathbf{x}_{i,k}^t}{\sum_{i=1}^n u_{l,i}^m} \quad (5)$$

The distance function between a sample and the static and temporal prototype of a cluster is computed following equation 6, where δ represents the euclidean distance. The membership degree of sample i to cluster l is demonstrated in equation 7.

$$d_\lambda^2(\mathbf{v}_l^s, V_l^t, x_i) = ||\mathbf{v}_l^s - \mathbf{x}_i^s||^2 + \lambda \sum_{k=1}^p \delta(\mathbf{v}_{l,k}^t, \mathbf{x}_{i,k}^t) \quad (6)$$

$$u_{l,i} = \frac{1}{\sum_{o=1}^c \left(\frac{d_\lambda(\mathbf{v}_l^s, V_l^t, x_i)}{d_\lambda(\mathbf{v}_o^s, V_o^t, x_i)} \right)^{\frac{2}{m-1}}} \quad (7)$$

Equation 8 presents the augmented FCM objective function.

$$J = \sum_{l=1}^c \sum_{i=1}^n u_{l,i}^m d_\lambda^2(\mathbf{v}_l^s, V_l^t, x_i) \quad (8)$$

The MFC algorithm is described in algorithm 1. Its inputs are the static X^s and temporal data X^t, number of clusters c, initial partition matrix U, fuzzification parameter m and temporal component weight λ which, in this study, was selected by grid search. It returns the final partition matrix U and the static V^s and temporal V^t prototypes.

Algorithm 1 Mixed Fuzzy Clustering (MFC)

Input:
X^s: $n \times r$ matrix of static data
X^t: $n \times q \times p$ matrix of temporal data
c: number of cluster prototypes
U: $c \times n$ initial partition matrix
m: fuzzification parameter
λ: temporal component weight
Output:
U: $c \times n$ partition matrix
V^s: $c \times r$ static cluster prototypes
V^t: $c \times q \times q$ temporal cluster prototypes
ΔJ: variation of objective function
while $\Delta J < \epsilon$ **do**
    Compute the static cluster prototypes V^s
    for k in $\{0, ..., p\}$ **do**
        Compute the temporal cluster prototype v_k^t
    end for
    Compute the distances d_λ
    Update the partition matrix U
    Compute ΔJ
end while

III. Takagi-Sugeno fuzzy modeling

Fuzzy models are "grey box" and transparent models that allow the approximation of non-linear systems with no previous knowledge of the system to be modeled. Fuzzy models have the advantage, in comparison to other non-linear modeling techniques, of not only providing transparency but also linguistic interpretation in the form of rules.

In this work, Takagi-Sugeno fuzzy models (TS-FM) [7] are derived from the data. These consist of fuzzy rules where each rule describes a local input-output relation. With TS-FM, each discriminant function consists, for the binary classification case, of rules of the type

$$R_j: \text{ If } x_1 \text{is } A_{j1} \text{ and } ... \text{ and } x_M \text{ is } A_{jM}$$
$$\text{then } d_j(\mathbf{x}) = f_j(\mathbf{x}), j = 1, 2, ..., K \quad (9)$$

where f_j is the consequent function of rule R_j and $M = r+q$, the number of features used. The output of the discriminant function $d_j(\mathbf{x})$ can be interpreted as a score (or evidence) for the the positive example given the input feature vector $\mathbf{x}$. The degree of activation of the jth rule is given by $\beta_j = \prod_{h=1}^M \mu_{A_{jh}}(\mathbf{x})$, where $\mu_{A_{jh}}(\mathbf{x}) : \mathbb{R} \to [0, 1]$. The discriminant output is computed by aggregating the individual rules contributions: $d(\mathbf{x}) = \frac{\sum_{j=1}^K \beta_j f_i(\mathbf{x})}{\sum_{j=1}^K \beta_j}$. A sample $\mathbf{x}$ is considered positive if the score is higher than a certain γ threshold $d_j(\mathbf{x}) > \gamma$.

The number of rules K and the antecedent fuzzy sets A_{jh} are determined by fuzzy clustering in the product space of the input variables. Section IV explains in detail the proposed methodology used to determine the cluster centers. The consequent functions $f_i(\mathbf{x})$ are linear functions determined by ordinary-least squares (OLS) in the space of the input and output variables.

IV. Modeling Based on MFC

Three different modeling approaches are used in order to determine the parameters of the used fuzzy model: (1) FCM TS-FM in which the antecedent fuzzy sets are determined by FCM clustering; (2) MFC TS-FM in which the antecedent fuzzy sets are determined by the presented MFC approach; (3) MFC-FCM TS-FM in which the features are transformed using MFC and the antecedent fuzzy sets are determined using FCM clustering.

A. FCM Fuzzy Model

In FCM FM, the time invariant features and samples of the time variant features are all equally as features and the

antecedent fuzzy sets are determined using the partition matrix generated by FCM. This is one of the most commonly used clustering method for the identification of TS-FM and has been used in multiple health care applications [8].

B. Proposed MFC fuzzy model

In MFC FM, the antecedent fuzzy sets are determined based on the partition matrix generated by the MFC algorithm.

This methodology was developed based on the belief that the identification of the fuzzy membership functions should be based on a non-conventional clustering algorithm in the presence of a mix of time variant and invariant features, variant features should not be directly mixed with time invariant features when calculating distances and different time variant features should be dealt with separately also.

C. Proposed FCM model with MFC feature transformation

In MFC-FCM FM, the time invariant and variant features are initially clustered and the generated partition matrix is used as the feature set for a FCM fuzzy model. In this case, the number of features after transformation is equal to the number of clusters specified for the MFC algorithm and the features are equal to the degree of membership of each sample to the mixed clusters.

This approach can be seen as a type of feature transformation method for which the resulting features represent the degree of membership of each point to the different clusters generated by the MFC algorithm.

D. Processing of data

Medical datasets are typically very heterogeneous [9], due to its multiple and sometimes dissimilar sources. Each patient has different periods of time staying in medical facilities, during which distinct variables are documented. In addition, equipment and human faults, as well as missing measurements are frequent. Particularly for retrospective evaluations, results quality relies heavily on the pre-processing of original data to handle problems associated with these datasets such as large amounts of missing data, uneven sampling times and existence of outliers.

Missing data is a common occurrence in ICU databases either due to intentional reasons, i.e. data is irrelevant for the clinical problem under consideration and thus is not recorded, or unintentional reasons, when some kind of intervention or activity renders the data useless. In this work, patients and variables were initially selected in order to minimize missing data and, when recoverable, missing data was imputed and outliers, identified using expert knowledge, were corrected using ZOH (zero order holder), while unrecoverable data resulted in patient discarding. In order to deal with differences in the frequency of collection, samples in both vasopressors datasets are aligned with heart rate since this is the variable most frequently measured. For a detailed description of this process please refer to [10].

Since statistical methods rely on measures that consider the spreading/dispersion of values and do not consider the nature of data, they often lead to the loss of important information. In the case of health care data it is important however to consider variable measurements distant from other observations, since they can represent sudden variations in a patients' physiological condition. In this work, data outliers were removed using expert knowledge, meaning that values outside acceptable physiological ranges were deleted.

V. Experimental results

A. Datasets

This paper used two de-identified publicly available ICU databases, MIMIC II and MEDAN. The MIMIC II (Multi-parameter Intelligent Monitoring for Intensive Care) database is an ICU database from the Beth Israel Deaconess Medical Center in the United States [11]. It contains demographics, medications, laboratory results and other clinical data from 32,535 patients collected over a seven-year period. Three distinct datasets were built for classification using clinical and demographic information of adult patients (>15 years old at time of admission, in a total of 24,580 patients). The MEDAN database [12] contains data from patients under abdominal septic shock registered from 1998 to 2002 by medical documentation staff at 71 ICUs in Germany. This dataset contains demographics and measurements of physiological variables from 410 patients, collected during their stay in the ICU.

Table I summarizes the characteristics of the datasets obtained for each one of the applications after pre-processing. The number of samples, percentage of samples of the positive class, time variant and invariant features and the classification label are presented.

1) MIMIC II - Readmissions:

The first dataset was used to develop a classification model for the prediction of early readmissions. Patients readmitted to the ICU within a period of 24–72 hours after discharge and patients who only experienced one ICU stay and did not die within one year after discharge are respectively labeled as class 1 and class 0. The time series representation of 7 variables collected during the last 10 hours of the patients' stay at the ICU were used. The selected variables were chosen based on previous studies that made use of the same database [13]. Age, gender, weight on admission, SAPS I and SOFA scores on admission were also collected for each patient and used as time invariant inputs. SAPS I (Simplified Acute Physiology Score) gives a measure of the severity of disease and SOFA (Sequential Organ Failure Assessment) is used to determine the extent of organ failure. The final dataset included 2653 patients, from which 199 were readmitted within 24–72 hours after discharge.

2) MIMIC II - Pneumonia and pancreatitis:

The second and third datasets were used to derive models to predict the necessity of the administration of vasopressors in ICU patients for two specific subsets of patients: patients suffering from pancreatitis and patients suffering from pneumonia. Given that these patients are usually treated differently in terms of medication and surgery procedures, the circumstances

TABLE I: Summary of datasets of the health care applications.

Dataset	Samples	Balance	Time variant features	Time invariant features	Label
MIMIC II - Readmissions	2653	7.5%	Creatinine, lactic acid, NBP mean, platelets, temperature, heart rate and SpO2 - oxygen saturation	Age, gender, weight, SAPS II and SOFA score on admission	Readmission of patient in a period after discharge
MIMIC II - Pneumonia	1323	38.3%	Lactic acid, WBC, Arterial PaCO2 and NBP	Age, gender, weight, SAPS II and SOFA score on admission	Administration of vasopressors
MIMIC II - Pancreatitis	378	18.0%	Sodium, BUN and WBC	Age, gender, weight, SAPS II and SOFA score on admission	Administration of vasopressors
MEDAN	100	44.0 %	Serum creatinine, serum calcium, aterial pCO2, pH, haematocrit, serum sodium, leukocytes, haemoglobin, central venous pressure, temperature, heart rate and systolic blood pressure	Age, weight	Mortality prediction

related to the initiation of vasopressors are presumably also distinct; therefore models are built for each dataset separately.

The final dataset consisted in a total of 1323 pneumonia patients and 378 pancreatitis with 4 and 3 temporal input variables respectively, sampled during the patients' stay in the ICU, and 5 demographics for each dataset. The time variant variables were chosen based on a previous study by [15] which used a previous version of MIMIC II to find the best predictors of the need of vasopressors using a combination of fuzzy modeling with bottom-up for feature selection. In order to timely predict the initiation of vasopressor administration, a window of 2 hours of data collected before the administration was neither used for modeling nor for validation of the models. The data was then resampled considering only 10 hours before the window with a sampling time of one hour. The output consists in a binary classification with positive value if the patient was on administered vasopressors and zero if not.

3) MEDAN:

This dataset was used to develop classification models for mortality prediction of patients under abdominal septic shock through physiological features uniformly sampled during the patients' stay at the ICU. The pre-processing of the original data performed in [16] was used, assuring data quality.

The most relevant features determined in [17] were used, resulting in a dataset comprising records of 12 time variant features measured over different periods of time, with a global sampling time of 24 hours. The time series resulting from these measurements were used as the time variant input, while the patients demographic information, age and weight, formed the time invariant input. Additionally to this data pre-processing and in order to maintain equal lengths of time series regarding each feature, only the last 10 days of patient care were considered, resulting in 10 sample points per feature. In this approach, patients with less than 10 measures per feature were disregarded. The final dataset comprises 100 patients, from which 44 did not survive (labeled as class 1).

B. Results

The three FM approaches presented were applied to each dataset, with the model parameters $\lambda = 0, ..., 2$ and $c = 2, ..., 4$ being selected by grid search. The datasets present distinct characteristics and challenges such as dimension and balance, shown on Table I. To overcome this, models were created using both the real distribution and by forcing a balanced distribution of train sets (with 25, 50 and 75% of samples belonging to the positive class). On both cases, the real distribution was used to evaluate the models on the test sets.

The performance of the models is evaluated in terms of area under the receiver operating characteristic curve (AUC) [18], accuracy (correct classification rate), sensitivity (true positive classification rate) and specificity (true negative classification rate). Cross validation is used to assess the validity and robustness of the models. Due to the diversity of dataset dimensions, this validation was performed using 5 and 10 folds. Table II displays the results for each FM approach, obtained using the best conditions for each dataset.

The results show that the models based on both MFC based approaches achieved better results than the FCM based models in three out of the four datasets used, pneumonia, pancreatitis and MEDAN. Furthermore, the only dataset in which the MFC could not improve the results, the readmissions dataset, had overall poorer results using either approach. This dataset is characterized by a highly imbalanced class distribution, which could not be entirely overcome even with the use of a balanced train set. It is also noticeable that the best method depends on the data tested. While for pancreatitis and pneumonia datasets the best results were achieved through the MFC FM, for the MEDAN dataset the MFC-FCM FM showed far better classification abilities, despite the methods poor results in all other datasets. Comparing the best conditions found for each problem, the results show that larger datasets, readmissions, pneumonia and pancreatitis achieved better performances based on the validation using 10 folds while for the smallest dataset, MEDAN, it was necessary to validate the results using only 5 folds, in order to increase the dimension of each test set and consequently provide a better evaluation of the models classification abilities.

In the presence of an highly imbalanced set, careful thought needs to be put due to the possibility of generating test sets with a reduced number of positive examples. For the most imbalanced datasets, readmissions and pancreatitis, a train set balanced with 50% positive class samples managed

TABLE II: Results of FM approaches for each dataset.

Dataset	Model	c	AUC	Accuracy	Sensitivity	Specificity
MIMIC II – Readmissions	FCM FM	2	0.58 ± 0.04	0.60 ± 0.03	0.55 ± 0.10	0.60 ± 0.05
	MFC-FM	2	0.58 ± 0.04	0.59 ± 0.03	0.56 ± 0.09	0.59 ± 0.05
	MFC-FCM FM	4	0.48 ± 0.07	0.62 ± 0.14	0.31 ± 0.17	0.65 ± 0.19
MIMIC II – Pneumonia	FCM FM	2	0.69 ± 0.04	0.71 ± 0.04	0.61 ± 0.05	0.77 ± 0.04
	MFC-FM	2	$0.70 \pm 0,04$	0.72 ± 0.04	0.63 ± 0.05	0.77 ± 0.04
	MFC-FCM FM	2	0.54 ± 0.04	0.53 ± 0.04	$0.57 \pm 0,07$	0.50 ± 0.06
MIMIC II – Pancreatitis	FCM FM	2	0.65 ± 0.07	0.68 ± 0.07	0.60 ± 0.17	0.70 ± 0.11
	MFC-FM	2	0.66 ± 0.08	0.69 ± 0.07	0.63 ± 0.20	0.70 ± 0.11
	MFC-FCM FM	2	0.47 ± 0.06	0.48 ± 0.07	0.46 ± 0.20	0.49 ± 0.11
MEDAN	FCM FM	2	0.54 ± 0.11	0.55 ± 0.11	0.48 ± 0.12	0.61 ± 0.18
	MFC-FM	3	0.60 ± 0.14	0.60 ± 0.14	0.53 ± 0.21	0.66 ± 0.21
	MFC-FCM FM	2	0.82 ± 0.07	0.82 ± 0.05	0.78 ± 0.26	0.85 ± 0.17

to improve the models performance. The pneumonia and MEDAN datasets did not require this balancing.

VI. Conclusions

This work proposes an efficient Mixed Fuzzy Clustering algorithm in order to handle datasets containing time variant and invariant features, converging their information to improve knowledge extraction. In addition, two different approaches are presented to derive Takagi-Sugeno fuzzy models based on this clustering algorithm. Their benefits are demonstrated by comparing the performance of the proposed methods to the modeling based on classic fuzzy clustering for different health care applications. Results show improvement in three out of four datasets.

Following this work, feature selection should be further researched for the proposed methods in order to determine which time variant and invariant features are best suited in this application and clustering algorithm.

Acknowledgment

This work was supported by FCT, through IDMEC, under project IC4U (PTDC/EMS-SIS/3220/2012). The work of J. L. Viegas was supported by the PhD in Industry Scholarship SFRH/BDE/95414/2013 from FCT and Novabase. S. Vieira acknowledges support by Program Investigador FCT (IF/00833/ 2014) from FCT, co-funded by the European Social Fund (ESF) through the Operational Program Human Potential (POPH).

References

[1] L. A. Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery. big data in the intensive care unit. closing the data loop. *American journal of respiratory and critical care medicine*, 187(11):1157–1160, 2013.

[2] S. M. Vieira, J. P. Carvalho, A. S. Fialho, S. R. Reti, S. N. Finkelstein, and J. M. C. Sousa. A decision support system for icu readmissions prevention. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*, pages 251–256. IEEE, 2013.

[3] T. Warren Liao. Clustering of time series data - A survey. *Pattern Recognition*, 38:1857–1874, 2005.

[4] S. Kisilevich, F. Mansmann, M. Nanni, and S. Rinzivillo. Spatio-temporal clustering. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 855–874. Springer US, 2010.

[5] H. Izakian, W. Pedrycz, and I. Jamal. Clustering spatiotemporal data: An augmented fuzzy C-means. *IEEE Transactions on Fuzzy Systems*, 21(5):855–868, 2013.

[6] J. C. Bezdek, R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984.

[7] T. Takagi and M. Sugeno. Fuzzy Identification of Systems and Its Application to Modeling and Control, 1985.

[8] A. L. Horn, F. Cismondi, A. S. Fialho, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D. Howell, and S. N. Finkelstein. Multi-objective performance evaluation using fuzzy criteria: Increasing sensitivity prediction for outcome of septic shock patients. In *Proceedings of 18th world congress of the international federation of automatic control (IFAC)*, volume 18, pages 14042–14047, 2011.

[9] J. Paetz, B. Arlt, K. Erz, K. Holzer, R. Brause, and E. Hanisch. Data quality aspects of a database for abdominal septic shock patients. *Computer Methods and Programs in Biomedicine*, 75:23–30, 2004.

[10] F. Cismondi, A. S Fialho, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D Howell, and S. N. Finkelstein. Computational intelligence methods for processing misaligned, unevenly sampled time series containing missing data. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, pages 224–231. IEEE, 2011.

[11] Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical care medicine*, 39(5):952–60, May 2011.

[12] E. Hanisch, R. Brause, B. Arlt, J. Paetz, and K. Holzer. The MEDAN Database, 2003.

[13] A. S. Fialho, F. Cismondi, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein. Data mining using clinical physiology at discharge to predict ICU readmissions. *Expert Systems with Applications*, 39(18):13158–13165, 2012.

[14] A. S. Fialho, L. A. Celi, F. Cismondi, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein. Disease-based modeling to predict fluid response in intensive care units. *Methods of Information in Medicine*, 52:494–502, 2013.

[15] A. S. Fialho, F. Cismondi, S. M. Vieira, J. M. C. Sousa, S. R. Reti, L. Celi, M. D. Howell, and S. N. Finkelstein. Fuzzy modeling to predict administration of vasopressors in intensive care unit patients. *IEEE International Conference on Fuzzy Systems*, (ii):2296–2303, 2011.

[16] F. J. Marques, A. Moutinho, S. M. Vieira, and J. M. C. Sousa. Preprocessing of clinical databases to improve classification accuracy of patient diagnosis. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 18:14121–14126, 2011.

[17] A. S. Fialho, F. Cismondi, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D. Howell, and S. N. Finkelstein. Predicting outcomes of septic shock patients using feature selection based on soft computing techniques. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, pages 65–74. Springer, 2010.

[18] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(4):29–36, 1982.