Chapter 4
Mixed Fuzzy Clustering for Deriving Predictive Models in Intensive Care Units

Cátia M. Salgado, Susana M. Vieira, and João M.C. Sousa

4.1 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 (Celi et al. 2013). Such systems acquire large quantities of data to discover hidden associations and understand patterns and trends in data, which can be used for diagnostic, prognostic and therapy (Celi et al. 2013; Vieira et al. 2013).

ICU databases contain records of patients' vital signs, laboratory results, prescribed and administrated medications, fluid balance, nursing notes, imaging reports, demographic information and clinical history. The premise of this work is that the complex and non linear relationship between different types of variables should be investigated and accounted for when deriving predictive models to support medical decision making. However, the mixed nature of data in electronic medical records still poses a daunting challenge for deriving data-based predictive models in the ICU. In particular, the available options to simultaneously handle static and time variant data are limited, and traditional methods do not provide the required means to extract useful knowledge that accounts for correlations between both.

In Izakian et al. (2013), the fuzzy c-means clustering technique is augmented for spatiotemporal clustering, a form of grouping objects based on their spatial (static) and temporal similarity. We extend this algorithm to any data set containing both time variant and time invariant features (mixed datasets) of any size, and use the knowledge extracted from the identified mixed structures to derive fuzzy models,

C.M. Salgado (✉) • S.M. Vieira • J.M.C. Sousa
IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, Lisboa, Portugal
e-mail: catia.salgado@tecnico.ulisboa.pt; susana.vieira@tecnico.ulisboa.pt; jmsousa@tecnico.ulisboa.pt

C. Kahraman, Y.I. Topcu (eds.), *Operations Research Applications in Health Care Management*, International Series in Operations Research & Management Science 262, https://doi.org/10.1007/978-3-319-65455-3_4

following the work of Ferreira et al. (2015). Fuzzy modeling can provide 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. We propose two different fuzzy modeling strategies based on the mixed fuzzy clustering algorithm and compare it with similar strategies based on fuzzy c-means. We show the results in four health care applications, for the prediction of adverse events of critically ill patients, including medication administration, mortality and readmissions.

The next section presents an overview of ICU adverse events. The mixed fuzzy clustering algorithm and the proposed modeling strategies are presented in Sects. 4.3 and 4.4, respectively. Data description and pre-processing is presented in Sect. 4.5, while experimental results for each application are shown in Sect. 4.6. Main conclusions and future work are presented in Sect. 4.7.

4.2 Adverse Events in the ICU

Several studies have focused on determining the predictive value of different types of numerical information in the prediction of adverse events, for improving the outcome of patients in the ICU. In particular, studies have focused on the prediction of mortality (Badawi and Breslow 2012; Clermont et al. 2001; Daly et al. 2001; Frize et al. 2001; Hug and Szolovits 2009; Ouanes et al. 2012; Reini et al. 2012; Ferreira et al. 2015), readmissions (Badawi and Breslow 2012; Campbell et al. 2008; Fialho et al. 2013, 2012; Frost et al. 2010; Gajic et al. 2008; Ouanes et al. 2012; Strand and Flaatten 2008; Walsh and Hripcsak 2014; Zheng et al. 2015; Ferreira et al. 2015) and length of stay (Frize et al. 2001; Reini et al. 2012).

Scoring systems in clinical use in the ICU include the Modified Early Warning Score (MEWS) (Reini et al. 2012), the Stability and Workload Index for Transfer (SWIFT) (Gajic et al. 2008), the Simplified Acute Physiology Score (SAPS II) (Le Gall et al. 1993), the Acute Physiology and Chronic Health Evaluation (APACHE II) (Knaus et al. 1985) and the Sequential Organ Failure Assessment (SOFA) (Vincent et al. 1996). MEWS attributes a score to each patient based on their respiratory, circulatory and neurological state, renal function and body temperature. SWIFT has been specifically developed to predict readmissions to the ICU; it represents a measure of five parameters: ICU admission source, length of ICU stay, respiratory impairment and neurological status, using the Glasgow Coma Scale (GCS) at discharge. APACHE II gives a score of the severity of illness; it is based on the patient's age and punctual physiological measurements of temperature, mean arterial pressure, pH, heart rate, respiratory rate, sodium, potassium, creatinine, hematocrit, white blood cells counts, GCS and partial pressure of arterial oxygen (PaO2), taken during the first 24 h after admission.

Currently, there has been an attempt to improve these conventional standard logistic regression techniques using machine learning algorithms such as artificial neural networks, fuzzy logic and decision trees, which resulted in a number of predictive models in different ICUs (Fialho et al. 2013, 2012; Walsh and Hripcsak 2014; Zheng et al. 2015; Salgado et al. 2016). In spite of the growing popularity of these models among the research community, the role they play in supporting the physicians' decisions and in improving patients' outcomes remains uncertain (Allaudeen et al. 2011; Kansagara et al. 2011; Ouanes et al. 2012).

To the best of the authors knowledge, the simultaneous mining of time series and time invariant data has not been addressed in any of the aforementioned studies, which suggests that information about the patients may be being lost in the data mining process. Hence, we developed predictive models based on MFC using information about variables with more than one record over the patient' stay, including vital signs and laboratory results, and information that remains static throughout the stay, such as gender, weight and admission status. Two main cohorts of patients are considered: patients in septic shock and patients that were readmitted within 24–72 h of discharge.

4.2.1 Septic Shock: Vasopressors Administration and Mortality

Sepsis is a systemic whole-body inflammatory response to infection. Septic shock is an outcome of a sepsis reaction, associated with multiorgan failure and out of the normal range measurements of blood pressure, temperature, respiratory rate and white blood cells counts. It is one of the most common reasons of death in intensive care units, with a mortality rate of about 50%. A patient is considered to be in septic shock when the hypotensive state related to a sepsis condition persists, causing severe malfunction of vital organs despite adequate fluid resuscitation.

The initial priority in managing septic shock is to restore blood pressure and cardiac output by intravenous fluids administration. When fluid resuscitation is unable to restore an adequate arterial pressure, therapy with vasopressors must be initiated. Vasopressors are hypotension blood vessels drugs that are very effective in increasing blood pressure. The procedure of vasopressors administration is risky, since the catheter insertion involves a surgical procedure that can be associated with infections. These complications are increased when the procedure is done urgently such as in the case of unexpected systemic shock. Knowing beforehand which patients will need vasopressors would reduce the number of times the procedure is implemented and reduce the associated risk of infection, since clinicians would have more time to safely initiate the central line insertion protocol. This would in turn substantially improve the outcomes of these patients.

When the septic shock is caused mainly by abdominal indications it is called abdominal septic shock. Previous works have applied knowledge-based neural networks (Paetz 2003) and neuro-fuzzy techniques for predicting the outcome of these patients. In Fialho et al. (2010), the authors implement ant colony and bottom-up tree search techniques, combined with fuzzy modeling and neural networks, to determine the set of features more correlated with the mortality of these patients.

This study attempts to predict vasopressors administration and mortality of ICU patients in abdominal septic shock.

4.2.2 Early Readmissions

Patients readmitted to the ICU have increased mortality, morbidity and length of stay. ICU readmissions are regarded as an indicator of poor care and represent increased costs to the hospital (Allaudeen et al. 2011; Boudesteijn et al. 2007). According to Elliott et al. (2014), ICU readmission rates reported in literature vary between 1.3 and 13.7%. Although patients can have an unplanned readmission for any reason, from incomplete treatment or poor care to poor coordination of services at the time of discharge and afterwards, many of these readmissions are potentially preventable (Goldfield et al. 2008). Adequate risk stratification at the time of discharge could reduce readmission rates and hence improve patient outcomes.

Risk factors for ICU readmission have been systematically reported in prospective and retrospective cohort studies. The most commonly identified factors include: patient location before ICU admission, SAPS II and APACHE II scores at admission, age, co-morbidities, ICU length of stay, physiologic abnormalities at the time of ICU discharge or on the ward, ICU discharge at night or after hours, discharge to another critical care area or hospital, shock index (heart rate/systolic blood pressure), respiratory rate, Glasgow Coma Score, and higher Nursing Activity Score at the time of discharge (Rosenberg and Watts 2000; Rosenberg et al. 2001).

4.3 Mixed Fuzzy Clustering

Mixed fuzzy clustering (MFC) is a novel clustering method based on fuzzy c-means (Bezdek et al. 1984) which deals with both time variant and time invariant features (Ferreira et al. 2015). 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. Each entity x_i, with $i = 1, \ldots, N$, is characterized by features that are constant during the sampling time in analysis, $\mathbf{x}_i^s$, and by features that change over time (multiple time series), X_i^t:

$$x_i = (\mathbf{x}_i^s, X_i^t) \tag{4.1}$$

The time invariant component of the entities is represented by $\mathbf{x}_i^s$, where R is the number of time invariant features:

$$\mathbf{x}_i^s = (x_{i1}^s, \ldots, x_{iR}^s) \tag{4.2}$$

In order to extend the spatiotemporal clustering method proposed in Izakian et al. (2013), 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. The time variant component of the entities is represented by the matrix X_i^t:

$$X_i^t = \begin{pmatrix} x_{i11}^t & x_{i12}^t & \cdots & x_{i1P}^t \\ x_{i21}^t & x_{i22}^t & \cdots & x_{i2P}^t \\ \vdots & \vdots & \ddots & \vdots \\ x_{iQ1}^t & x_{iQ2}^t & \cdots & x_{iQP}^t \end{pmatrix}, \tag{4.3}$$

where Q is the length of time series.

The time invariant prototypes $\mathbf{v}_l^s$ for each cluster l are given by:

$$\mathbf{v}_l^s = \frac{\sum_{i=1}^N u_{li}^m \mathbf{x}_i^s}{\sum_{i=1}^N u_{li}^m}, \tag{4.4}$$

where $l = 1, \ldots, C$. The time variant prototypes $\mathbf{v}_{lk}^t$ for each cluster l and feature k are given by:

$$\mathbf{v}_{lk}^t = \frac{\sum_{i=1}^N u_{li}^m \mathbf{x}_{ik}^t}{\sum_{i=1}^N u_{li}^m} \tag{4.5}$$

The matrix of time variant prototypes for cluster l is represented by V_l^t.

In the above equations, the membership degree u_{li} of entity i to cluster l is given by:

$$u_{li} = \frac{1}{\sum_{o=1}^C \left(\frac{d_\lambda^2(\mathbf{v}_l^s, V_l^t, x_i)}{d_\lambda^2(\mathbf{v}_o^s, V_o^t, x_i)} \right)^{\frac{1}{m-1}}}, \tag{4.6}$$

where $m \in]1, \infty]$ is a weighting exponent that controls the degree of fuzziness.

The MFC algorithm clusters the data 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. The distance function between a sample and the time invariant and time variant prototype of a cluster is computed:

$$d_\lambda^2(\mathbf{v}_l^s, V_l^t, x_i) = ||\mathbf{v}_l^s - \mathbf{x}_i^s||^2 + \lambda \sum_{k=1}^P ||\mathbf{v}_{lk}^t - \mathbf{x}_{ik}^t||^2 \tag{4.7}$$

The augmented FCM objective function is given by:

$$J = \sum_{l=1}^C \sum_{i=1}^N u_{li}^m d_\lambda^2(\mathbf{v}_l^s, V_l^t, x_i) \tag{4.8}$$

The MFC is described in Algorithm 1. Its inputs are the time invariant $\mathbf{x}^s$ and time variant data X^t, number of clusters C, initial partition matrix $U = [u_{li}]$, degree of fuzziness m and time variant component weight λ. It returns the final partition matrix $U = [u_{li}]$ and the time invariant V^s and time variant V^t prototypes. J^n represents the objective function in iteration n.

Algorithm 1 Mixed fuzzy clustering (MFC)

```
**Input:**
$C$: number of cluster prototypes
$m$: degree of fuzziness
$\lambda$: time variant component weight
$\mathbf{x}^s$: $N \times R$ matrix of time invariant data
$X^t$: $N \times Q \times P$ matrix of time variant data
$\mathbf{U}$: $C \times N$ random initial partition matrix
$\epsilon$: stop criterion
**Output:**
$\mathbf{U}$: $C \times N$ partition matrix
$\mathbf{V}^s$: $C \times R$ matrix of time invariant cluster prototypes
$V^t$: $C \times Q \times P$ matrix of time variant cluster prototypes
**while** $\Delta J > \epsilon$ **do**
    Compute the cluster prototypes $\mathbf{v}_l^s$
    **for** $k$ in $\{1, \ldots, P\}$ **do**
        Compute the cluster prototypes $\mathbf{v}_{lk}^t$
    **end for**
    Compute the distances $d_\lambda^2(\mathbf{v}_l^s, V_l^t, x_i)$
    Update the partition matrix $\mathbf{U}$
    Compute $\Delta J = J^n - J^{n-1}$
    $n = n + 1$
**end while**
```

4.4 Fuzzy Modeling Based on Mixed Fuzzy Clustering

4.4.1 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 of not only providing transparency, but also linguistic interpretation in the form of rules.

In this work, Takagi-Sugeno (TS) fuzzy models (FMs) (Takagi and Sugeno 1985) are derived from data. These consist of fuzzy rules describing a local input-output relation. With TS FM, each discriminant function consists of rules of the type:

$$\begin{aligned} R_i &: \text{If } x_1 \text{ is } A_{i1} \text{ and } \ldots \text{ and } x_M \text{ is } A_{iM} \\ &\text{then } y(\mathbf{x}) = f_i(\mathbf{x}), \quad i = 1, 2, \ldots, C, \end{aligned} \tag{4.9}$$

where A_{ij} are the antecedent fuzzy sets, f_i is the consequent function of rule R_i and y is the output. The degree of activation of the ith rule is given by $\beta_i = \prod_{j=1}^{M} \mu_{A_{ij}}(\mathbf{x})$, where $\mu_{A_{ij}}(\mathbf{x}) : \mathbb{R} \to [0, 1]$. The output is computed by aggregating the individual rules contributions:

$$y(\mathbf{x}) = \frac{\sum_{i=1}^{C} \beta_i f_i(\mathbf{x})}{\sum_{i=1}^{C} \beta_i} \tag{4.10}$$

The number of rules C and the antecedent fuzzy sets A_{ij} are determined by fuzzy clustering in the space of the input and output variables. The consequent functions $f_i(\mathbf{x})$ are linear functions determined by ordinary least squares.

For classification, a threshold γ is required to turn the continuous output $y \in [0, 1]$ into the binary output $y \in \{0, 1\}$. This way, an entity $\mathbf{x}$ is labeled as 1 if $y(\mathbf{x}) \geqslant \gamma$ and as 0 otherwise.

4.4.2 Proposed TS Fuzzy Models

Distinct Takagi Sugeno fuzzy model approaches based on clustering were considered for this study. The strategies differ in the type of input data and in the methodology used to determine the antecedent fuzzy sets. In particular, the antecedent fuzzy sets and the number of rules of the TS fuzzy model are determined based either on the partition matrix generated by the FCM algorithm (FCM FM), or in the partition matrix generated by MFC (MFC FM). MFC FM 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 time invariant features. Time variant features should not be directly mixed with time invariant features when calculating distances, and different time variant features should also be dealt with separately.

The input variables consist of (1) time variant and time invariant features or (2) transpose of the partition matrix generated by MFC ($\mathrm{U}^{\mathrm{MFC}}$). When time variant and time invariant data are used as input for the FCM FM, each time stamp of the time series is treated as one feature, i.e., the input of the model consists of a $N \times (R + Q \times P)$ matrix. When using the partition matrix, each feature corresponds to the degree of membership of the entities to the clusters such that the number of features equals the number of clusters determined in the clustering step. In particular, the input becomes u_{il}, where $i = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, C$, which corresponds to the transpose of the partition matrix in (4.6). This approach can be seen as a type of feature transformation method.

Two approaches based on MFC are presented: MFC FM and FCM–$\mathrm{U}^{\mathrm{MFC}}$ FM, and compare it with the traditional FCM FM. The approaches listed below are described in Algorithm 2.

- FCM FM: Antecedent fuzzy sets determined by FCM in the space of the input and output variables.
- MFC FM: Antecedent fuzzy sets determined by MFC in the space of the input and output variables.
- FCM–U^{MFC} FM: Antecedent fuzzy sets determined by FCM in the space of the partition matrix generated by MFC and output variable.

4.5 Data Description

This paper uses two de-identified publicly available ICU databases, MIMIC II and MEDAN, which are described in the following.

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 (Saeed et al. 2011). Version 2.6 used in this study contains demographics, medications, laboratory results and other clinical data from 32,535 patients, collected over a 7-year period. Three datasets were built for classification using clinical, demographic and score information of adult patients (>15 years old at time of admission).

The MEDAN database (Hanisch et al. 2003) 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.

4.5.1 Data Processing

Medical datasets are typically very heterogeneous (Paetz et al. 2004), 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 seldom measurements are frequent. Particularly for retrospective evaluations, the quality of results relies heavily on the preprocessing of original data. Problems commonly associated with these datasets are the existence of missing data, uneven sampling times and 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. When recoverable, missing data was filled using the zero order hold procedure, while unrecoverable data led to patient discarding.

Algorithm 2 Takagi-Sugeno fuzzy models

```
Data:
$\mathbf{x}^s$: $N \times R$ matrix of time invariant data
$X^t$: $N \times Q \times P$ matrix of time variant data
$\mathbf{Y}$: $N \times 1$ vector of class labels
$[\mathbf{x}^s \parallel X^t]$: matrix of concatenated input data
$[\mathbf{x}^s \parallel X^t \parallel \mathbf{Y}]$: matrix of concatenated input and output data
Parameters:
$C$: number of cluster prototypes
$\mathbf{U}$: $C \times N$ initial partition matrix
$m$: degree of fuzziness
$\lambda$: time variant component weight
$\mathrm{U}^{\mathrm{MFC}}$: $C \times N$ final MFC partition matrix
$(\mathrm{U}^{\mathrm{MFC}})^T$: $N \times C$ transpose matrix of $\mathrm{U}^{\mathrm{MFC}}$
$\mathrm{U}^{\mathrm{FCM}}$: $C \times N$ final FCM partition matrix
function $trainFCM$:($[\mathbf{x}^s \parallel X^t \parallel \mathbf{Y}]$, $C$, $m$)
    train a TS FM using product-space FCM clustering
    return model
end function
function $trainMFC$:($[\mathbf{x}^s \parallel X^t \parallel \mathbf{Y}]$, $C$, $m$, $\lambda$, $U^{\mathrm{MFC}}$)
    train a TS FM using product-space MFC clustering
    return model
end function
procedure FCM FM
    $[\mathrm{U}^{\mathrm{FCM}}, V] = \mathrm{FCM}([\mathbf{x}^s \parallel X^t \parallel \mathbf{Y}], C, \mathbf{U}, m)$
    model=trainFCM($[\mathbf{x}^s \parallel X^t \parallel \mathbf{Y}]$, $C$, $m$)
    $Y_m$=test($[\mathbf{x}^s \parallel X^t]$, model)
end procedure
procedure MFC FM
    $[\mathrm{U}^{\mathrm{MFC}}, \mathbf{v}_l^s, V_l^t] = \mathrm{MFC}([\mathbf{x}^s \parallel \mathbf{Y}], X^t, C, \mathbf{U}, m, \lambda)$
    model=trainMFC($[\mathbf{x}^s \parallel X^t \parallel \mathbf{Y}]$, $C$, $m$, $\mathrm{U}^{\mathrm{MFC}}$)
    $Y_m$=test($[\mathbf{x}^s \parallel X^t]$, model)
end procedure
procedure FCM–$\mathrm{U}^{\mathrm{MFC}}$ FM
    $[\mathrm{U}^{\mathrm{MFC}}, \mathbf{v}_l^s, V_l^t] = \mathrm{MFC}([\mathbf{x}^s \parallel \mathbf{Y}], X^t, C, \mathbf{U}, m, \lambda)$
    model=trainFCM($[(\mathrm{U}^{\mathrm{MFC}})^T \parallel \mathbf{Y}]$, $C$, $m$)
    Compute distance from $\mathbf{x}^s$ to $\mathbf{v}_l^s$ and $X^t$ to $V_l^t$
    Update the partition matrix $\mathrm{U}^{\mathrm{MFC}}$
    $Y_m$=test($\mathrm{U}^{\mathrm{MFC}}$, model)
end procedure
```

Since statistical methods rely on measures that consider the spreading 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 patient physiological condition. In this work, data outliers were removed using expert knowledge, meaning that values outside the acceptable physiological ranges were deleted.

4.5.2 Vasopressors Administration

MIMIC II database was used to derive models to predict the necessity of vasopressors administration in 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 related to the initiation of vasopressors are also presumably distinct; therefore models are built for each dataset separately.

The datasets contain data regarding: patients that received one of the following vasopressors: levophed, dopamine, epinephrine, vasopressin and neosynephrine; the patients' first ICU stay (in order to consider the first administration of vasopressors); a period of at least 6 h between the initiation of data acquisition and the administration of vasopressors; patients that were on vasopressors for more than 2 h.

The final pancreatitis and pneumonia datasets contain 378 and 1323 patients, respectively. Time variant features were sampled during the length of stay of the patient in the ICU, whereas time invariant features were selected from demographic information and scores records on admission. Time variant variables were selected based on a previous study (Fialho et al. 2011), where the best predictors of the need of vasopressors where determined using a combination of fuzzy modeling and bottom-up for feature selection. In order to predict in a timely manner the initiation of the vasopressor administration, a window of 2 h of data collected before the administration was not used neither for modeling nor for validation of the models. The data was then resampled considering only 10 h before the window with a sampling time of 1 h. The output consists of a binary classification with positive value if the patient was on vasopressors.

List of physiological variables, demographics and score records extracted from the MIMIC II database for vasopressors administration classification in pancreatitis and pneumonia patients:

Time variant (Pancreatitis):
Sodium (mEq/L)
BUN: Blood urea nitrogen (mg/dL)
WBC: White blood cells ($\times 10^3$cells/μL)
Time variant (Pneumonia):
Lactic acid (mg/dL)
WBC: White blood cells ($\times 10^3$cells/μL)
PaCO2: Arterial carbon dioxide partial pressure (mmHg)
NBP: Non-invasive blood pressure mean (mmHg)
Time invariant (Pneumonia and Pancreatitis):
Age (years)
Weight (kg)
SAPS II on admission: Simplified Acute Physiology Score
SOFA on admission: Sequential Organ Failure Assessment

4.5.3 *Mortality in Abdominal Septic Shock*

The MEDAN database was used to develop classification models for mortality prediction of patients under abdominal septic shock. The pre-processing of the original data performed by Marques et al. (2011) was used, assuring data quality.

The most relevant features determined by Fialho et al. (2010) were selected, resulting in a dataset containing records of 12 time variant features measured over different periods of time, with a global sampling time of 24 h. The time series resulting from these measurements were used as the time variant input, while patients' demographic information, age and weight, formed the time invariant input. 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 discarded. The final dataset comprises 100 patients, from which 44 did not survive (labeled as class 1).
List of physiological variables and demographic information extracted from the MEDAN database for mortality classification in abdominal septic shock patients:

Time variant:
Arterial pCO2 (mmHg)
Central venous pressure (cmH2O)
Hematocrit (%)
Hemoglobin (g/dl)
Heart rate (beats/min)
WBC (cells$\times 10^3/\mu$L)
pH
Serum calcium (mmol/L)
Serum creatinine (mg/dL)
Serum sodium (mmol/L)
Systolic blood pressure (mmHg)
Temperature (°C)
Time invariant:
Age (years)
Weight (kg)

4.5.4 *Readmissions*

MIMIC II was used to develop models for the prediction of early readmissions. Time variant data consists of the time series representing the last 10 measurements of 7 variables, collected during the patients' stay at the ICU. The selected variables were chosen based on previous studies that made use of the same database (Fialho et al. 2012). Age, weight on admission, SAPS II and SOFA scores on admission were also collected for each patient and used as time invariant inputs. Patients readmitted to the ICU within a period of 24–72 h after discharge and patients who only experienced one ICU stay and did not die within 1 year after discharge were respectively labeled as class 1 and class 0. The final dataset includes 2653 patients, from which 199 were readmitted within 24–72 h after discharge.

List physiological variables and demographic information extracted from the MIMIC II database for readmissions classification:

Time variant:
Creatinine (mg/dL)
Lactic acid (mg/dL)
NBP: Non-invasive blood pressure mean (mmHg)
Platelets (×10^3 cells/μL)
Temperature (°C)
Heart rate (beats per minute)
SpO2: Oxygen saturation in the blood (%)
Time invariant:
Age (years)
Weight (kg)
SAPS II on admission: Simplified Acute Physiology Score
SOFA on admission: Sequential Organ Failure Assessment

4.6 Results

The performance of the models is evaluated in terms of area under the receiver operating characteristic curve (AUC) (Hanley and McNeil 1982), accuracy (correct classification rate), sensitivity (true positive classification rate) and specificity (true negative classification rate).

The datasets are evaluated using fivefold cross validation. For each fold, a grid search is performed on the training set to find the λ that maximizes the AUC; the training set is randomly divided into two smaller subsets *T1* and *T2*, where *T1* represents 50% of the training set. *T1* is used for training and *T2* for tuning. For this model, a range of values of the threshold γ are evaluated on set training set and the γ that results in the smallest difference between sensitivity and specificity is used. The model is tested using data the model has not yet used for training or tuning. For FCM–U$^{\mathrm{MFC}}$, the input U used for testing the models is obtained by computing the partition matrix using the cluster prototypes obtained in training. For updating the matrix, the distance $d^2_\lambda(\mathbf{v}^s_l, V^t_l, x_i)$ between the test entities and the cluster prototypes is computed. Cross validation is performed separately for $C = \{2, 3, \ldots, 10\}$, and results are averaged over the folds. In the end, results are shown for the C giving the best average.

Figure 4.1 shows the average performance in terms of AUC, for each C, whereas the best average results are shown in Table 4.1.

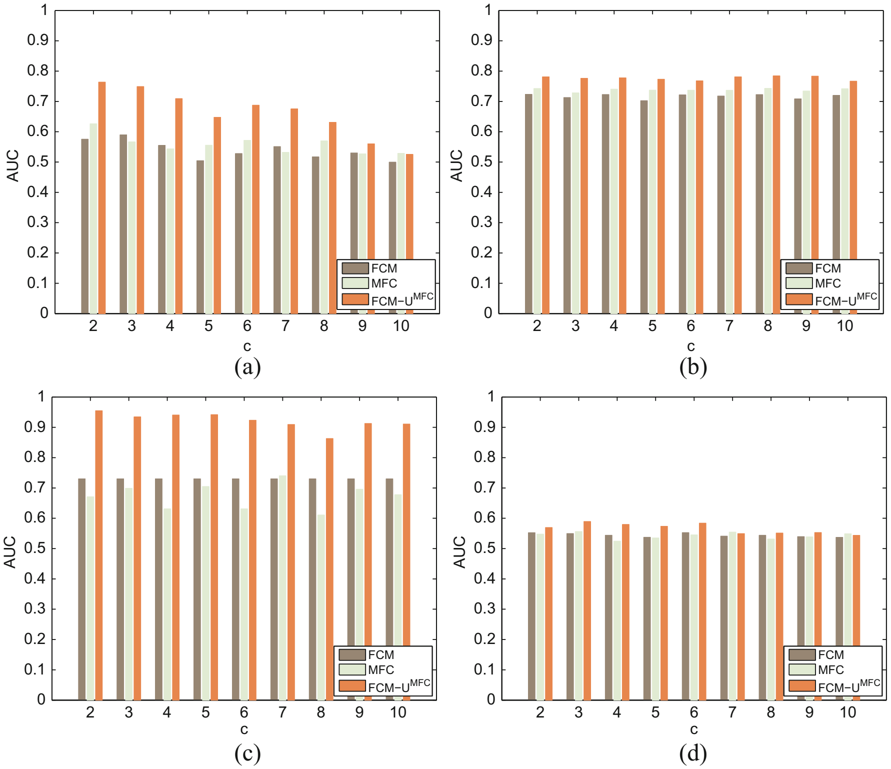

Fig. 4.1 Performance of FCM, MFC and FCM–U$^{\text{MFC}}$ fuzzy models for different number of clusters, for MEDAN, readmissions, pancreatitis and pneumonia datasets. (**a**) Pancreatitis. (**b**) Pneumonia. (**c**) MEDAN. (**d**) Readmissions

In order to investigate the influence of λ in the performance of each method, boxplots showing the selected values of λ for different number of clusters are presented in Fig. 4.2. Overall, the choice of λ seems to be greatly affected by the data divisions, exception made to the MFC method in MEDAN and pneumonia datasets.

FCM–U$^{\text{MFC}}$ FM performs better than the other methods in all datasets. In particular, FCM–U$^{\text{MFC}}$ FM increases the AUC of MEDAN and pancreatitis by a factor of approximately 30%, when compared to FCM FM. MFC FM has also improved the performance of FCM FM in all datasets, due to its ability of creating rules that adapt to the mixed nature of data. Table 4.1 and Fig. 4.2 show that for all health care applications investigated, the information contained in the time variant variables is more relevant in predicting the output than the information contained in the time invariant variables.

Compared to a previous study using the same datasets (Ferreira et al. 2015), there is an overall increase in the performance of all FMs. Two main reasons can be

Table 4.1 Results with fivefold cross validation

Dataset	Fuzzy models	C	AUC	ACC	Sensitivity	Specificity
MEDAN	FCM	10	0.73 ± 0.09	0.64 ± 0.08	0.62 ± 0.14	0.66 ± 0.11
	MFC	7	0.74 ± 0.11	0.66 ± 0.09	0.61 ± 0.12	0.70 ± 0.11
	FCM–U$^{\text{MFC}}$	2	**0.96 ± 0.05**	**0.86 ± 0.05**	**0.91 ± 0.09**	**0.82 ± 0.13**
Readmissions	FCM	6	0.55 ± 0.03	0.53 ± 0.04	0.54 ± 0.05	0.53 ± 0.05
	MFC	3	0.56 ± 0.06	0.53 ± 0.04	0.56 ± 0.07	0.53 ± 0.03
	FCM–U$^{\text{MFC}}$	3	**0.59 ± 0.02**	**0.55 ± 0.02**	**0.58 ± 0.02**	**0.55 ± 0.02**
Pancreatitis	FCM	3	0.59 ± 0.08	0.57 ± 0.08	0.57 ± 0.14	0.56 ± 0.10
	MFC	2	0.63 ± 0.13	0.58 ± 0.12	**0.68 ± 0.20**	0.54 ± 0.13
	FCM–U$^{\text{MFC}}$	2	**0.76 ± 0.05**	**0.70 ± 0.07**	0.65 ± 0.03	**0.72 ± 0.08**
Pneumonia	FCM	2	0.72 ± 0.05	0.66 ± 0.04	0.69 ± 0.07	0.64 ± 0.04
	MFC	8	0.74 ± 0.05	0.68 ± 0.06	**0.72 ± 0.05**	0.66 ± 0.08
	FCM–U$^{\text{MFC}}$	8	**0.78 ± 0.04**	**0.70 ± 0.02**	0.70 ± 0.09	**0.70 ± 0.04**

attributed to this fact: first, model parameters λ and C accept a wider range of values, and second, in MIMIC II datasets, the gender is no longer used as a time invariant input variable, which means that the finding of structures based on the distribution of classes is not hampered by another binary variable.

Previous studies, namely Fialho et al. (2011, 2013), achieved notable results in the prediction of continuous vasopressors administration. However, these studies do not provide the means for predicting the initiation of therapy. Hence the comparison of performance results in Table 4.1 should not be straightforward.

Next follows a more detailed discussion of the results obtained for each dataset.

4.6.1 Vasopressors Administration

In the pancreatitis dataset, the AUC of FCM–U$^{\text{MFC}}$ FMs decreases with increasing number of clusters (see Fig. 4.1a). Results suggest that patients in this group are better divided in two or three subgroups. In the case of FCM–U$^{\text{MFC}}$, λ values are close to 1, whereas for MFC, λ oscillates between 0.5 and 6, as shown in Fig. 4.2a. When the number of subgroups increases, a higher weight is in general given to the time variant part of data.

Contrarily to the previous vasopressors dataset, increasing number of clusters produce little changes in the AUC of pneumonia models (see Fig. 4.1b), showing that for this particular case, increasing the number of rules does not provide an added value in the prediction of the output. Thus, $C = 2$ should be considered for all FM approaches, with λ between 1.5 and 3.

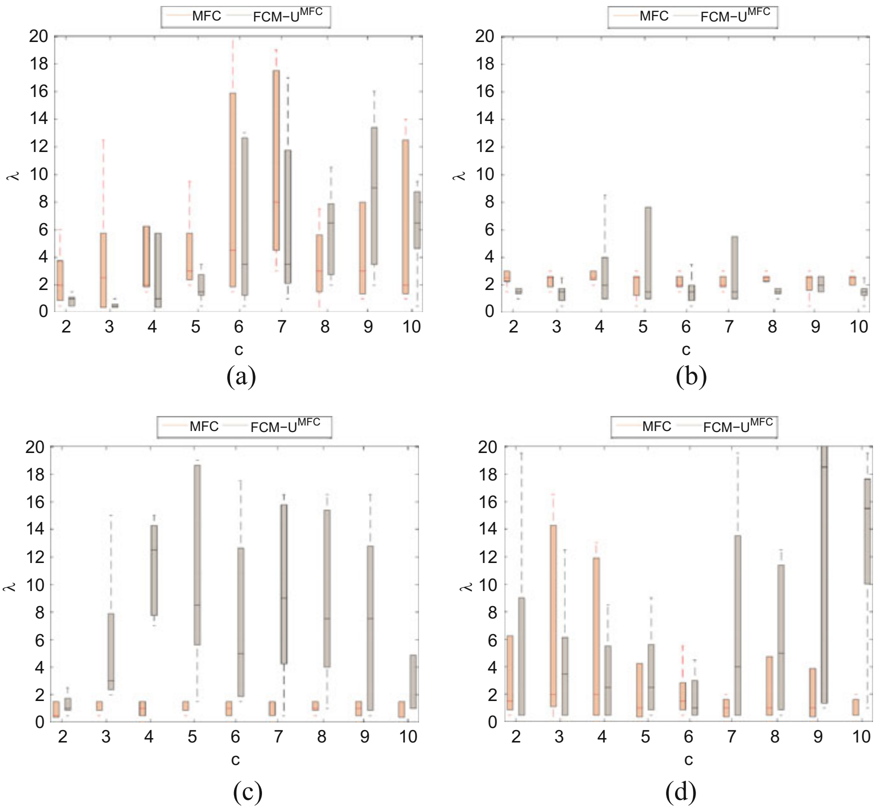

Fig. 4.2 Boxplots of λ associated to different number of clusters C, for MEDAN, readmissions, pancreatitis and pneumonia datasets. (**a**) Pancreatitis. (**b**) Pneumonia. (**c**) MEDAN. (**d**) Readmissions

4.6.2 *Mortality Prediction*

FCM–U$^{\text{MFC}}$ tends to select higher values of λ, except when $C = 2$, while MFC tends to select values close to 1, i.e, it gives the same weight to both time variant and time invariant components of data, as shown in Fig. 4.2c. The fact that λ is different between the rounds, oscillating between 0 and 2, justifies the differences between MFC and FCM FM approaches, highlighting the importance of this parameter in the tuning of the models. If λ would equal 1 in all rounds of MFC, results were expected to deviate less and be more similar between MFC and FCM approaches. The fact that higher values of λ are associated with improved performance also highlights the importance of this parameter.

Figure 4.1c shows that for both FCM and MFC FM, the C selected by grid search is not the best option. Smaller values of C achieve nearly the same performance, at lower computational costs and simplified model interpretability.

Thus, for this dataset, 2 clusters and 3 clusters (or rules) would be sufficient to derive FCM and MFC models, respectively. Nonetheless, the best strategy is still to perform dimensionality reduction by transforming the input variables into degrees of membership to 2 clusters.

While FCM and MFC FM approaches perform similarly to Fialho et al. (2010) when using a reduced number of features, FCM–U^{MFC} significantly improved previous results (AUC$=0.75 \pm 0.01$ vs AUC$=0.96 \pm 0.05$).

4.6.3 Readmissions

For this dataset, varying number of clusters result in small changes in the performance, and high values of λ are in general selected for both MFC approaches, as shown in Figs. 4.1d and 4.2d.

Monitoring signals, such as NBP mean, temperature, heart rate and SpO2 are associated with higher sampling rates than laboratory results such as lactic acid, platelets and creatinine. Having this in mind, the fact that all time series have a length of 10 points means that the data used for modeling contains laboratory measurements that can go up to 10 days of each patient' stay (when applicable, otherwise missing data is filled using the ZOH), and measurements of monitoring signals of the last 10 h of stay, approximately. We point this misalignment as the main probable reason for the overall poor results in comparison to approaches using the mean values during the last 24 h (Fialho et al. 2012). Hence, further studies should be conducted in order to handle unevenly and misaligned time series. Other reasons may be pointed out to justify the poor results, namely the highly imbalanced class distribution.

4.7 Conclusions

This work presents two modeling strategies based on the mixed fuzzy clustering algorithm, in order to handle datasets containing time variant and time invariant features, converging their information to improve knowledge extraction. One strategy uses Takagi-Sugeno where the antecedent fuzzy sets are determined by MFC in the product space of the time variant and time invariant variables and the other strategy uses Takagi-Sugeno where the antecedents are determined based on FCM in the product space of the membership degrees derived by MFC.

The performance of models is tested in four health care datasets, for the classification of critically ill patients, and is compared with Takagi-Sugeno based on FCM. The best method, FCM–U^{MFC}, is common to all applications: mortality in abdominal septic shock patients is classified with an AUC of 0.96 ± 0.05, readmissions to the ICU with 0.59 ± 0.02 and vasopressors administration in pancreatitis and in pneumonia patients with 0.76 ± 0.05 and 0.78 ± 0.04, respectively. The findings of this work suggest that dimensionality reduction based on the transformation of input variables into degrees of membership allows the finding of important structures in data, hence the discover of relevant rules in the knowledge base system.

Future work should focus on finding which time variant and time invariant features are best predictors of different adverse events in the ICU, using wrapper feature selection methods for performance improvement.

Acknowledgements This work is funded by Portuguese Funds through the FCT, Foundation for Science and Technology under the project UID/EMS/50022/2013 and IF/00833/2014/CP1238/CT0002. S.M. Vieira acknowledges the 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

Allaudeen N, Schnipper JL, Orav EJ, Wachter RM, Vidyarthi AR (2011) Inability of providers to predict unplanned readmissions. J Gen Intern Med 26(7):771–776

Badawi O, Breslow MJ (2012) Readmissions and death after ICU discharge: development and validation of two predictive models. PloS One 7(11):e48758

Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10(2):191–203

Boudesteijn E, Arbous S, van den Berg P (2007) Predictors of intensive care unit readmission within 48 hours after discharge. Crit Care 11(Suppl 2):P475

Campbell AJ, Cook JA, Adey G, Cuthbertson BH (2008) Predicting death and readmission after intensive care discharge. Br J Anaesth 100(5):656–662

Celi LA, Mark RG, Stone DJ, Montgomery RA (2013) "big data" in the intensive care unit. closing the data loop. Am J Respir Crit Care Med 187(11):1157–1160

Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT (2001) Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 29(2):291–296

Daly K, Beale R, Chang R (2001) Reduction in mortality after inappropriate early discharge from intensive care unit: logistic regression triage model. BMJ 322(7297):1274

Elliott M, Worrall-Carter L, Page K (2014) Intensive care readmission: a contemporary review of the literature. Intensive Crit Care Nurs 30(3):121–137

Ferreira MC, Salgado CM, Viegas JL, Schäfer H, Azevedo CS, Vieira SM, Sousa JMC (2015) Fuzzy modeling based on mixed fuzzy clustering for health care applications. In: 2015 IEEE international conference on FUZZ-IEEE

Fialho AS, Cismondi F, Vieira SM, Sousa JMC, Reti SR, Howell MD, Finkelstein SN (2010) 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. Springer, Berlin, pp 65–74

Fialho AS, Cismondi F, Vieira SM, Sousa JMC, Reti SR, Celi LA, Howell MD, Finkelstein SN (2011) Fuzzy modeling to predict administration of vasopressors in intensive care unit patients. IEEE Int Conf Fuzzy Syst (ii):2296–2303

Fialho AS, Cismondi F, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2012) Data mining using clinical physiology at discharge to predict ICU readmissions. Expert Syst Appl 39(18):13158–13165

Fialho AS, Celi LA, Cismondi F, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2013) Disease-based modeling to predict fluid response in intensive care units. Methods Inf Med 52:494–502

Frize M, Ennett CM, Stevenson M, Trigg HC (2001) Clinical decision support systems for intensive care units: using artificial neural networks. Med Eng Phys 23(3):217–225

Frost SA, Tam V, Alexandrou E, Hunt L, Salamonson Y, Davidson PM, Parr MJ, Hillman KM (2010) Readmission to intensive care: development of a nomogram for individualising risk. Crit Care Resusc 12(2):83–89

Gajic O, Malinchoc M, Comfere TB, Harris MR, Achouiti A, Yilmaz M, Schultz MJ, Hubmayr RD, Afessa B, Farmer JC (2008) The stability and workload index for transfer score predicts unplanned intensive care unit patient readmission: Initial development and validation*. Crit Care Med 36(3):676–682

Goldfield NI, McCullough EC, Hughes JS, Tang AM, Eastman B, Rawlins LK, Averill RF (2008) Identifying potentially preventable readmissions. Health Care Financ Rev 30(1):75

Hanisch E, Brause R, Arlt B, Paetz J, Holzer K (2003) The MEDAN Database. http://www.medan.de.

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(4):29–36

Hug CW, Szolovits P (2009) ICU acuity: real-time models versus daily models. In: Association AMI (ed) AMIA annual symposium proceedings, vol 2009, p 260

Izakian H, Pedrycz W, Jamal I (2013) Clustering spatiotemporal data: an augmented fuzzy C-means. IEEE Trans Fuzzy Syst 21(5):855–868

Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S (2011) Risk prediction models for hospital readmission: a systematic review. J Am Med Assoc 306(15):1688–1698

Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) Apache II: a severity of disease classification system. Crit Care Med 13(10):818–829

Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (saps II) based on a European/North American multicenter study. J Am Med Assoc 270(24):2957–2963

Marques FJ, Moutinho A, Vieira SM, Sousa JMC (2011) Preprocessing of clinical databases to improve classification accuracy of patient diagnosis. IFAC Proc Vol (IFAC-PapersOnline) 18:14121–14126

Ouanes I, Schwebel C, Français A, Bruel C, Philippart F, Vesin A, Soufir L, Adrie C, Garrouste-Orgeas M, Timsit J, Misset B (2012) A model to predict short-term death or readmission after intensive care unit discharge. J Crit Care 27(4):422–e1

Paetz J (2003) Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. Artif Intell Med 28(2):207–230

Paetz J, Arlt B, Erz K, Holzer K, Brause R, Hanisch E (2004) Data quality aspects of a database for abdominal septic shock patients. Comput Methods Prog Biomed 75:23–30

Reini K, Fredrikson M, Oscarsson A (2012) The prognostic value of the modified early warning score in critically ill patients: a prospective, observational study. Eur J Anaesthesiol (EJA) 29(3):152–157

Rosenberg AL, Watts C (2000) Patients readmitted to ICUs*: a systematic review of risk factors and outcomes. CHEST J 118(2):492–502

Rosenberg AL, Hofer TP, Hayward RA, Strachan C, Watts CM (2001) Who bounces back? Physiologic and other predictors of intensive care unit readmission. Crit Care Med 29(3):511–518

Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39(5):952

Salgado CM, Vieira SM, Mendonça LF, Finkelstein S, Sousa JMC (2016) Ensemble fuzzy models in personalized medicine: application to vasopressors administration. Eng Appl Artif Intell 49:141–148

Strand K, Flaatten H (2008) Severity scoring in the ICU: a review. Acta Anaesthesiol Scand 52(4):467–78

Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern 15(1):116–132

Vieira SM, Carvalho JP, Fialho AS, Reti SR, Finkelstein SN, Sousa JMC (2013) A decision support system for ICU readmissions prevention. In: IFSA world congress and NAFIPS annual meeting (IFSA/NAFIPS), 2013 joint. IEEE, Piscataway, NJ, pp 251–256

Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG (1996) The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 22(7):707–710

Walsh C, Hripcsak G (2014) The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. J Biomed Inform 52:418–426

Zheng B, Zhang J, Yoon SW, Lam SS, Khasawneh M, Poranki S (2015) Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Syst Appl 42(20):7110–7120