Linguistic Summaries of Intensive Care Unit Septic Shock Patients

Rui Jorge Almeida

***Abstract*— Linguistic summarization is a data mining and knowledge discovery approach to extract patterns and sum up large volume of data into simple sentences. There is a large research in generating linguistic summaries which can be used to better understand and communicate about patterns, evolution and long trends in numerical, time series or labelled data. The objective of this work is to develop a computational system capable of automatically generate linguistic descriptions in time series data containing labelled data, not only of the whole series, but also on the differences between subsets of the data. For this purpose we propose a new type of differential summaries, based on a numerical criterion assessing the behaviour of the summary on each subset of interest. Furthermore, this paper proposes an extension of linguistic summaries to provide temporal and categorical contextualisation. This is of particular interest in healthcare to detect differences related to a condition or illness as well as the effectiveness of the administered treatment.**

I. Introduction

THE rapid progress of information technology has facilitated the availability of huge amounts of data. Analysis of these huge data and their non-trivial trends may be complicated. Data mining or knowledge discovery methods to automatically summarize the data and reveal trends or non-trivial dependencies are highly desirable. Linguistic summaries (LS) are examples of such methods, that produce concise, human-consistent description of a data set [1]. This concept was extended and further developed by Kacprzyk and Yager [2] and by Kacprzyk, Yager and Zadrożny [3]. According to this approach, numerical data can be summarized and presented in the form of natural language like sentences, called protoforms, as e.g., “Most senior workers have high salary”, which are interpreted using the framework of Zadeh’s [4] calculus of linguistically quantified propositions.

Various techniques to develop linguistic summaries in an automatic manner can be found in the literature, and generally speaking follow two distinct paths [5], one using natural language generation and the other using fuzzy logic tools. In this research we focus on the latter. Linguistic summaries are usually modelled using type-1 fuzzy sets, but type-2 fuzzy sets can also be used [6], [7]. Many authors generate linguistic summaries using protoforms, such as “most employees are young” [8], [9], [10], [11], but recently it has been proposed to perform linguistic summaries in databases using If–Then rules [7]. These If–Then rules provide a linguistic description of the database and can also be used for prediction.

Rui Jorge Almeida are with Erasmus University Rotterdam, Erasmus School of Economics, Department of Econometrics, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (email: rjalmeida@ese.eur.nl).

The author would like Bernadette Bouchon-Meunier, Marie-Jeanne Lesot, Gilles Moyse and Uzay Kaymak for the support and feedback that has contributed to the successful completion of this research.

This work was supported by the IEEE-CIS Walter Karplus Student Research Grant.

Most applications of linguistic summaries have been on the business field (see *e.g.* [8], [9], [12]), but many studies dealing with healthcare [13], [11], [10] also exist. A comparisson between the similarities of a set of linguistic summaries in different time periods for different investment funds are studied in [12]. It is also possible to compare time series based on the result of user defined queries over a data cube with time dimension. The similarity between time series is then described using local changes [14]. In [8] linguistic summaries of investment funds are obtained using a set of features to characterize the trends such as the slope of the line segment and study the description of duration and variability. A similar idea is used in [11] to provide summaries of changes in behaviour for elders, while [10] provides activity summaries for eldercare based on a 3D silhouette representation of an elder is presented in [13]. The issue of continuous monitoring of eldercare, received further attention in [15], [16], [17], [18] with different approaches to compute distance between linguistic summaries to define the presence of abnormal conditions and aggregate these linguistic summaries.

The objective of this research is to obtain descriptive models of events to aid decision making. The dataset under study is composed of intensive care unit abdominal septic shock patients. This serious condition is not fully understood and differences between patients are not easily identified. This work focuses on generating not only a general summary for all patients but also in highlighting the differences exhibited in patients with different class labels. For this we propose to extend the protoforms of linguistic summaries as defined in [1] to provide categorical contextualization. From these summaries it is possible to clearly identify differences between categories. To that aim, we propose a new type of summaries, defined as differential, based on a numerical criterion to compare linguistic summaries. This criterion is used to differentiate each subset of the data identified by a category label. Furthermore, since the data set under study is composed of observations on multiple phenomena observed over multiple time periods for the same objects, we propose linguistic summaries that provide temporal contextualization, explicitly quantifying attributes and time.

The outline of the paper is as follows. In Section II, we provide the basic approach to linguistic summarization of databases and a dissimilarity metric between linguistic summaries. In Section III we propose an extension of linguistic summaries to explicitly consider objects and their time context. In Section IV we further extend linguistic summaries to include categorical label, from where we can obtain a novel type of differential linguistic summaries, based on a dissimilarity metric to highlight differences between linguistic summaries of objects with different category labels. An example of the proposed summaries applied to patients with abdominal septic shock is presented in Section V. Finally, conclusions and future work are given in Section VI

II. Related Works

A. Linguistic Summaries as Protoforms

In this section we briefly present the basic approach to linguistic summarization of databases as defined by [1] and extended in [2], [3]. From this approach we propose extensions to include category and temporal contextualisation, presented in the following sections.

1) Linguistic Expression and Components: Given a finite set of objects $Y = \{y_1, \dots, y_n\}$ in a database D and a set of attributes $A = \{A_1, \dots, A_p\}$ describing objects from Y, classic protoforms to define a linguistic summary depend on two components, a summarizer P, a quantifier Q, and possibly on an additional qualifier R, taking one of the following forms

$$Q\, y' s \text{ are } P \tag{1}$$

$$QR\, y' s \text{ are } P\,. \tag{2}$$

An example of (1) is “Most patients are tall” and of (2) is “Most young patients are tall”.

More formally, the summarizer P is a set of w fuzzy modalities F_{A_j}, $j = 1..w$, with $w \leq p$, associated to data attributes (*e.g.* the modality low defined differently for the attributes blood pressure and heart rate). It is modelled using:

$$\mu_P = \mu_{F_{A_1}} \wedge \ldots \wedge \mu_{F_{A_w}} \tag{3}$$

where $\wedge$ is a t-norm.

The quantifier Q is a linguistic quantifier (*e.g.* most) measuring the agreement in quantity, associated to a membership function μ_Q. The qualifier R is another attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_k determining a (fuzzy) subset of Y (*e.g.* young for attribute age).

A measure of validity or truth T is associated with this representation, it is a number from the interval $[0, 1]$ assessing the truth of the summary. It can be calculated using Zadeh’s [4] calculus of quantified propositions. This measure determines the degree to which a linguistically quantified proposition equated with a linguistic summary is true. For the linguistic summary (1), this measure is defined as:

$$T = \mu_Q \left(\frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right) \tag{4}$$

$$T = \mu_Q \left(\frac{\sum_{i=1}^{n} \mu_P(y_i) \wedge \mu_R(y_i)}{\sum_{i=1}^{n} \mu_R(y_i)} \right) \tag{5}$$

respectively.

B. Similarity between Linguistic Summaries

In this section we briefly describe the distance metric between summaries based on fuzzy protoforms presented in [15], closely following their notations. This dissimilarity measure takes into account not only the linguistic meaning of the summaries, but also numeric characteristic attached to them, such their truth values and their degrees of focus, as defined below.

Given two linguistic summaries $LS_1 = Q_1 R_1\, y' s \text{ are } P_1$ and $LS_2 = Q_2 R_2\, y' s \text{ are } P_2$ with truth values T_1 and T_2 respectively, the similarity is defined as [15]:

$$\begin{aligned} \text{sim}(LS_1, LS_2) = \min\big(&\text{sim}(P_1, P_2), \text{sim}(Q_1, Q_2), \\ &\text{sim}(R_1, R_2), \text{sim}(T_1, T_2)\big) \end{aligned} \tag{6}$$

where each individual similarity is detailed below. The induced dissimilarity

$$\begin{aligned} d(LS_1, LS_2) =& 1 - \text{sim}(LS_1, LS_2) \\ =& \max\big(1 - \text{sim}(P_1, P_2), 1 - \text{sim}(Q_1, Q_2), \\ & 1 - \text{sim}(R_1, R_2), 1 - \text{sim}(T_1, T_2)\big) \end{aligned} \tag{7}$$

is a metric on the space of protoform summaries [15].

The similarity between summarizers P_1 and P_2 depends whether the summarizers describe the same attributes or not and is calculated using

$$\text{sim}(P_1, P_2) = \min\left(\frac{a}{b}, \frac{\int (\mu_{P_1} \cap \mu_{P_2})}{\int (\mu_{P_1} \cup \mu_{P_2})} \right) \tag{8}$$

where a and b are respectively the number of common attributes for summarizers P_1 and P_2 and the total number of attributes involved in their union. For the case of a summarizer composed of several attributes, their cylindrical extension is used. Fractions a/b and $\int (\mu_{P_1} \cap \mu_{P_2}) / \int (\mu_{P_1} \cup \mu_{P_2})$ are Jaccard measures [15].

The similarity between quantifiers Q_1 and Q_2 is computed with the Jaccard measure

$$\text{sim}(Q_1, Q_2) = \frac{\int (\mu_{Q_1} \cap \mu_{Q_2})}{\int (\mu_{Q_1} \cup \mu_{Q_2})} \tag{9}$$

The similarity between qualifiers R_1 and R_2 is defined as

$$\begin{aligned} \text{sim}(R_1, R_2) = \min\Bigg(& \frac{\int (\mu_{R_1} \cap \mu_{R_2})}{\int (\mu_{R_1} \cup \mu_{R_2})}, \\ & 1 - \big|d_{\text{foc}}(R_1) - d_{\text{foc}}(R_2)\big| \Bigg) \end{aligned} \tag{10}$$

where $|\cdot|$ is the absolute value and d_{foc} is the degree of focus whose definition is recalled hereafter. If the protoforms are simple, *i.e.* R is absent, $\text{sim}(R_1, R_2) = 1$ which indicates that R is treated as being a fuzzy set that characterizes the whole universe Y.

The degree of focus limits the search for the best linguistic summaries and is defined as [9]:

$$d_{\text{foc}} = \frac{1}{n} \sum_{i=1}^{n} \mu_R(y_i)\,. \tag{11}$$

The essence of the degree of focus is to give the proportion of objects satisfying property R among all objects. The extended protoform linguistic summaries (2) limits the search space to a limited subspace of objects that fulfil an additional condition specified by qualifier R. If the degree of focus is high, then such a summary concerns many objects, that it is a general summary. The degree of focus can only be calculated for summaries of extended protoforms (2). It is fixed to the value 1 for simple protoforms (1).

Lastly, the similarity of truth values T_1 and T_2 can be calculated as

$$\mathrm{sim}(T_1, T_2) = 1 - |T_1 - T_2| \,. \tag{12}$$

III. Temporal Contextualisation

Databases are often composed of observations on multiple phenomena observed over long time periods for the same objects under study. In statistics and econometrics these databases are usually referred to as panel data. Depending on the type of study, the interest may lie in characterizing time series using local changes [14] or study the description of duration and variability of different trends [8]. In our approach we are interested in providing temporal contextualisation when summarizing objects with different attributes.

A. Proposed Protoforms

In this work we focus on explicitly characterizing attributes over time, to obtain summaries such as "Most patients have high blood pressure most of the time". We propose to extend the original protoforms (1) and (2):

$$Q\, y's \text{ are } P\, Q_T \text{ times} \tag{13}$$

$$QR\, y's \text{ are } P\, Q_T \text{ times} \tag{14}$$

where Q_T is a time quantifier.

We note that the linguistic interpretation of this type of linguistic summaries has a very different linguistic interpretation if the quantifier order is reversed, *i.e.* Q_T time, $Q\, y's$ are P. In this case, it is less clear the characterization of the attribute over time. We believe that the LS given by (13) and (14) are simpler to be human interpretable and can provide an adequate temporal contextualisation of events.

B. Proposed Evaluation of the Truth Degree

In order to assess the validity of the temporal summaries (13), we propose to compute their truth value as

$$T = \mu_Q \left(\frac{1}{n} \sum_{i=1}^{n} \mu_{Q_T} \left(\frac{1}{T} \sum_{t=1}^{T} \mu_P(y_{it}) \right) \right) \,, \tag{15}$$

where y_{it} indicates that the attribute under consideration evolves over time t for object y_i. To keep the calculation of the truth value consistent with the linguistic interpretation, this quality measure has to be calculated in the following order: for each object of the data base y_{it}, for a summarizer P we first process a given attribute A_j using a fuzzy predicate F_{A_j}. We then quantify the number of times that fuzzy predicate F_{A_j} occurs. The last necessary step is to quantify the number of objects that have the same (fuzzy) quantity of fuzzy predicate F_{A_j}. To clarify this, we will make use of an example. Let us assume that we are interested in the LS "Few patients have low heart rate most of the time". For each patient, first we fuzzify attribute heart rate, followed by fuzzifing the temporal quantity (number of occurrences) that heart rate is low. Finally we quantify how many patients have low Heart Rate most of the time.

For (14), we propose to extend of the previous truth value (15) in the same way as (5) extends (4):

$$T = \mu_Q \left(\frac{1}{T} \sum_{i=1}^{n} \mu_{Q_T} \left(\frac{\sum_{t=1}^{T} \mu_P(y_{it}) \wedge \mu_R(y_{it})}{\sum_{t=1}^{T} \mu_R(y_{it})} \right) \right) \,. \tag{16}$$

IV. Categorical Contextualisation

Considering the case of data for which category information is available, we also propose a categorical extension of linguistic summaries: we propose to use crisp category labels $C = \{c_1, \ldots, c_k\}$, as a form to provide insights into differences between patients or events in medical data. The category labels C refer to information contained in the data, such as a medical condition (*e.g.* disease), medical procedure (*e.g.* intubation) or status (*e.g.* deceased). This categorical data is crisp in nature.

A. Proposed Protoforms

A protoform of the form (2) could be used, by replacing R with $c \in C$. The reason for proposing a new protoform is to keep the idea of the original qualifier intact (*i.e.* another attribute together with a linguistic value), and maintain consistency with other quality measures [19]. We note that replacing R by c could be misleading because $\mu_c(y_i)$ would not refer to a linguistic value (in the form of a fuzzy predicate), but instead to crisp category data. Thus we propose to extend simple and complex protoforms in the form:

$$Q\, y's \text{ with } c \text{ are } P \tag{17}$$

$$QR\, y's \text{ with } c \text{ are } P \tag{18}$$

An example of this type of summaries would be "Most patients with disease X have low blood pressure". In this protoform the inclusion of crisp information in a linguistic summary is clear. This type of linguistic summaries also allows the use of indirect information and uses it as class labels. For example, patients with measurements of oxygen flow indicates that they are intubated.

The protoforms (17) and (18) can be further extended to provide both temporal and categorical contextualisation:

$$Q\, y's \text{ with } c \text{ are } P\, Q_T \text{ time} \tag{19}$$

$$QR\, y's \text{ with } c \text{ are } P\, Q_T \text{ time} \tag{20}$$

An example of these summaries is "Most patients with disease X have a low heart rate most of the time".

B. Proposed Evaluation of the Truth Degree

In calculations of quality measures, such as the truth value, for these linguistic summaries we are only interested in objects y_i which belong to a given class c. We define for any c category label a subset of Y as $Y^c = \{y_i \in Y / y_i \in c\}$ and the number of elements of this set is n^c. Naturally $Y = Y^{c_1} \cup Y^{c_2} \cup \ldots \cup Y^{c_k}$. The truth value for (17), (18), (19) and (20) can be obtained by substituting y_i for y_i^c and n for n^c in (1), (2), (13) and (14) respectively. For example the truth value for (17) is defined as

$$T = \mu_Q \left(\frac{1}{n^c} \sum_{y_i \in Y^c} \mu_P(y_i) \right) \; . \tag{21}$$

C. Differential Linguistic Summaries

1) Linguistic expression: In this work the focus is on generating linguistic descriptions in time series data, not only of the whole series, but also on the differences between subsets of the data identified with category labels. The objective is to characterise the category labels through the identifications of summaries that exclusively apply for one category label, but not others: the aim is to distinguish between the case where both summaries "most male patients have high heart rate" and "most female patients have high heart rate" are valid from the case where only one of them applies. In the latter case, we propose to underline the specifics of a category label by the definition of a differential linguistic summary, of the form "most male patients have high heart rate while female patients do not".

The proposed enriched linguistic summaries are composed of two parts: a part highlighting the differences between subsets of the data with different category labels and a part which refers to all category labels combined. The associated protoform is

Differences:

$$Q\, y's \,\text{with}\, c_1 \text{ are } P$$
$$\text{while}\, y's \,\text{with}\, c_2 \text{ do not.}\, (d, T) \tag{22}$$

Global:

$$Q\, y's \,(\text{with both } c_1 \& c_2) \text{ are } P.\, (T) \tag{23}$$

The first part (22) highlights differences between summaries with different classes. It is associated with two assessment criteria: d indicates the extent to which the summary indeed differentiates the two category labels, as detailed in the next subsection; T is the truth degree of the linguistic summary "$Q\, y's \,\text{with}\, C_1 \text{ are } P$". Two parameters are used to select the summaries to be part of the global summary: only summaries with high differential property $d \geq \alpha_1$ and high truth value $T \geq \alpha_2$ are kept. α_1 and α_2 are user-set parameters.

The second part (23) is composed of the general linguistic summaries where there is no large difference between linguistic summaries of different classes, but there is a high truth value. For the whole summary, only linguistic summaries with a truth value above a threshold α_3 are reported. It can be noted that this parameter can be set to the same value as α_2 or to a lower value to be less severe for summaries applying to all data. For this work we used a value of $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$.

The enriched linguistic summaries presented in (22) and (23) are illustrated for protoform (17). The differential summaries can also be based on the more complex form (18).

2) Evaluation: The evaluation of differential summaries is based on truth degrees and the differential criterion. Truth degrees are computed as presented in the previous subsection, see Equation (21). The aim of the differential criterion is to assess the extent to which a linguistic summary indeed characterises a categorical label, *i.e.* applies to it but not to others.

This criterion compares $LS_1 = Q\, y's \,\text{with}\, c_1 \text{ are } P$ and $LS_2 = Q\, y's \,\text{with}\, c_2 \text{ are } P$, *i.e.* two summaries with the same quantifier, temporal quantifier, summarizer and qualifier but different category labels. One of them must have a high truth degree and the other one a low truth degree. We therefore propose to simply define the differential criterion as

$$d = |T_1 - T_2| \, . \tag{24}$$

Using this definition, the negation "c_2 do not" in the differential summary "$Q\, y's \,\text{with}\, c_1 \text{ are } P \text{ while } y's \,\text{with}\, c_2$ do not" refers to the whole summary "$Q\, y's \text{ are } P$" and not only to the quantifier or summarizer.

A more general case can be considered, where two summaries slightly differing by their quantifier or summarizer are opposed one to another to define the differential summary, *i.e.* $LS_1 = Q_1\, y's \,\text{with}\, c_1 \text{ are } P_1$ and $LS_2 = Q_2\, y's \,\text{with}\, c_2 \text{ are } P_2$ with similar P_1 and P_2 or with similar Q_1 and Q_2. In the general case, we propose to define

$$d = d(LS1, LS2)\text{cmp}(c_1, c_2) \tag{25}$$

where $d(LS1, LS2)$ is the dissimilarity measure (7) applied to the linguistic summaries ignoring the category labels and $\text{cmp}(c_1, c_2)$ is a comparison measure for category labels defined as

$$\text{cmp}(c_1, c_2) = \begin{cases} 1 & \text{if } c_1 \neq c_2 \\ 0 & \text{otherwise} \end{cases} \; . \tag{26}$$

Using this definition, if the considered summaries $LS1$ $LS2$ apply to the same category label, they are associated to $\text{cmp}(c_1, c_2) = 0$ and thus to $d = 0$ and they do not satisfy the condition on minimal differential criterion. In the case where they are identical except for their categorical labels, as considered above, $\text{sim}(P_1, P_2) = 1$, $\text{sim}(Q_1, Q_2) = 1$, $\text{sim}(R_1, R_2) = 1$, $\text{sim}(Q_{T1}, Q_{T2}) = 1$ and $\text{cmp}(C_1, C_2) = 1$ (25) reduces to:

$$\begin{aligned} d =& 1 - \text{sim}(LS1, LS2) \\ =& 1 - \min\big(1, 1, 1, \text{sim}(T_1, T_2), 1, 1\big) \\ =& |T_1 - T_2| \; . \end{aligned} \tag{27}$$

which corresponds to (24).

V. Application to Medical Data

The considered application aims at generating informative linguistic descriptions of medical patients. Besides providing a general summary for all patients, we are also interested in highlighting the differences exhibited between patients with different class labels. This type of linguistic summaries can provide the decision maker with a comprehensive, human consistent summary of important differences and changes over long periods of time. The data set under study is composed of observations on multiple phenomena observed over multiple time periods for the same patients. We use new linguistic summaries that explicitly quantify objects and time. The differences between patient state, may be the result of a condition, illness or administered treatment. Since medical data sets are large, it is very difficult for a human being to capture, process and understand all changes.

A. Considered Data

This study used data from the MEDAN database [20]. This is a database composed of intensive care unit (ICU) abdominal septic shock patients admitted to 70 different hospitals in Germany, collected from 1998 to 2002. All information is anonymous. In this study we performed our experiments in a subgroup of 383 patients that meet the criteria for abdominal septic shock, and focused exclusively on physiological parameters, commonly assessed within the ICU setting. From this set of variables, we focused on a smaller subset of relevant variables [21]. All chosen variables are independent with minimal correlation. The primary outcome variable is the patient condition (alive or deceased) in a 24-hour window from a given time point. This variable is encoded in a binary format, taking value one if the patient died within that period of time, and zero if not.

As with other real-world databases, preprocessing of the data is necessary to improve its quality to be processed into linguistic summaries. In order to deal with variables collected with different sampling periods, a template variable is used. This process allows all variable samples to be available at the same point in time as the template variable. The template variable chosen is the heart rate, since it is the most frequently measured variable (in average one sample every 60 minutes) and thus, the one introducing fewer artefacts in the data [22].

For the case under study, ICU data can be missing either because because exogenous interventions or endogenous activities have rendered the data useless or they are perceived to be irrelevant for the current clinical problems [23]. When it is possible to prove that a variable was not measured during a certain period of time because of an intentional reason (e.g. ventilator parameters when a patient is extubated), this missing segment is considered as non-recoverable [22]. In this work, these non-recoverable missing segments are deleted. We note that this indirect information could also be used as class labels for the protoforms (17). On the other hand, if the variable is supposed to exist, but for some unintentional reason (*e.g.* sensor malfunction) it is missing, this absent segment is considered recoverable and thus, proper imputation techniques can be applied [22]. Missing values are not a problem for the derivation of linguistic summaries or the calculation of quality measures. Nonetheless, they may bias the quality measures. For example it is possible, that measurements were more frequent on time periods where the patient was exhibiting a higher heart rate, since probably he was deemed to be at risk. In this work, following the assumption that there are no huge variations between measurements, the last available value is used to impute values to these recoverable missing segments.

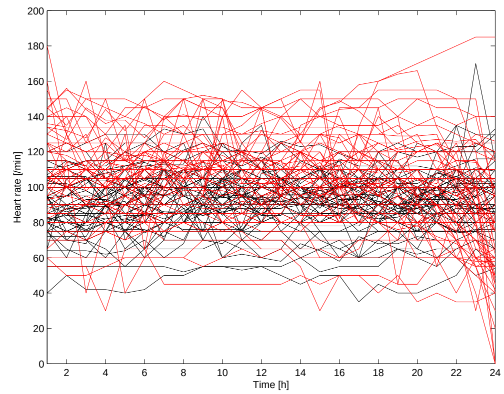

Fig. 1. Heart rate for septic shock patients. Patient state, black=alive, red=deceased

B. Categorical Summaries

In this section we provide linguistic summaries of patients observations of heart rate (HR) and also heart rate combined with values for the partial thromboplastin time (PTT) blood test. By using the protoform (22) we are able to differentiate patients with different category labels. This methodology is applied for all patient observations, *i.e.* we use all collected data for all patients combined. We refer to them in the summaries simply as observations. Although very simple because all patients observations are combined, these linguistic summaries provide a general overview of the differences between all measured observations of heart rate. Figure 1 shows the data for all patients under study. By observing this figure, we can see that there is no clear separation between patients with a different condition after the considered period. A way to identify possible differences aids practitioners' decision making by providing human interpretable summaries. These differences can also help to identify which measurements show large differences between patients.

1) Simple Summaries: We start by using the simplest protoforms (17) to summarize the patients observations of heart rate. We used fuzzy trapezoids to model the modalities of each attribute and quantifier. The fuzzy predicates used for the summarizer heart rate are shown in Fig. 2, while Fig. 3 shows the linguistic quantifier.

Differences:

- *few* observations of alive patients have *low* value of HR

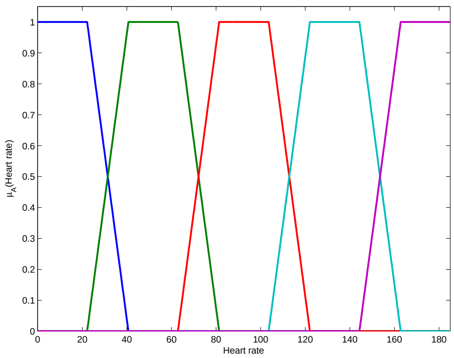

Fig. 2. Membership function for the summarizer heart rate. Left to right *'very low','low','medium','high','very high'*

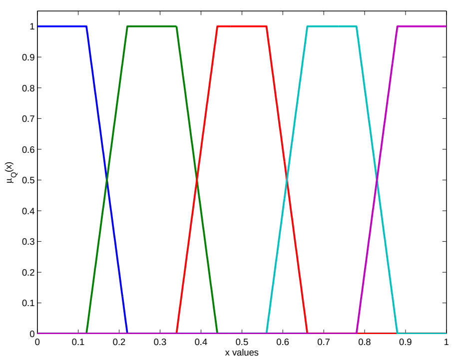

Fig. 3. Membership function for the quantifier Q. Left to right *'very few','few','half','most','almost all'*.

while deceased patients do not. ($d(LS_1, LS_2)$=1, T=1)

- *most* observations of alive patients have *medium* value of HR while deceased patients do not. ($d(LS_1, LS_2)$=0.51, T=0.51)
- *very few* observations of alive patients have *high* value of HR while deceased patients do not. ($d(LS_1, LS_2)$=1, T=1)
- *very few* observations of deceased patients have *low* value of HR while alive patients do not ($d(LS_1, LS_2)$=1, T=1)
- *few* observations of deceased patients have *high* value of HR while alive patients do not ($d(LS_1, LS_2)$=1, T=1)
- *half of the* observations of deceased patients have *medium* value of HR while alive patients do not ($d(LS_1, LS_2)$=0.51, T=1)

Global:

- *very few* observations have a *very low* value of HR. (T=1)
- *very few* observations have a *low* value of HR. (T=1)
- *most* observations have a *medium* value of HR. (T=1)
- *very few* observations have a *very high* value of HR. (T=1)
- *few* observations have a *high* value of HR. (T=0.83)

From the global summaries, we can observe that most observations have a medium value of heart rate. Interestingly, by observing the difference summaries, it is possible to

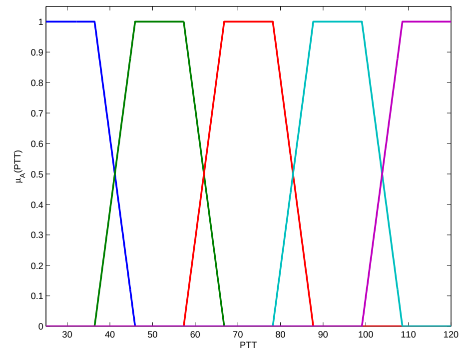

Fig. 4. Membership function for the summarizer PTT. Left to right *'very low','low','medium','high','very high'*

observe that this is also the case for observations of patients who were alive after the considered period, while for the patients who deceased this was only the case for half of them. For high values of heart rate, there are only very few of the observation of patients who lived, while there are more patients who deceased that exhibit high values of heart rate.

It can be noted that due to the difference between the total number of patient observations n and the number of patient observations with a given class n_C it is possible that a linguistic summary with the same quantifier and summarizer appear in both the differences and global part of the summary.

2) Extended Summaries: For linguistic summaries of data set containing several attributes, a simple approach to limit the number of summarizers is to limit the class of possible summaries by predefining a smaller subset of descriptors (*e.g.* low Heart Rate for patients) [19]. Since in most observations medium values of heart rate are observed, we use the extended linguistic summaries given by (18) to summarize the relation between patients observations of PTT with medium heart rate. These summaries also highlight the differences between patients with different classes. The fuzzy predicates medium used for the qualifier heart rate and summarizer PTT are shown in Fig. 2 and Fig. 4, respectively. Figure 3 shows the linguistic quantifier.

Differences:

- *most* observations of alive patients with *medium* value of HR also have a *very low* value of PTT while deceased patients do not. ($d(LS_1, LS_2)$=1, T=1)
- *few* observations of alive patients with *medium* value of HR also have a *low* value of PTT while deceased patients do not. ($d(LS_1, LS_2)$=0.95, T=1)
- *very few* observations of alive patients with *medium* value of HR also have a *medium* value of PTT while deceased patients do not. ($d(LS_1, LS_2)$=0.60,T=1)
- *half of the* observations of deceased patients with *medium* value of HR also have a *very low* value of PTT while alive patients do not. ($d(LS_1, LS_2)$=0.64, T=0.64)
- *half of the* observations of deceased patients with *medium* value of HR also have a *low* value of PTT while

alive patients do not. ($d(LS_1, LS_2)$=0.95, T=0.95)

- *few* observations of deceased patients with *medium* value of HR also have a *medium* value of PTT while alive patients do not. ($d(LS_1, LS_2)$=0.60, T=0.60)

Global:

- *half of the* observations with *medium* value of HR also have a *very low* value of PTT. (T=0.57).
- *few* observations with *medium* value of HR also have a *low* value of PTT. (T=1).
- *very few* observations with *medium* value of HR also have a *medium* value of PTT. (T=1).
- *very few* observations with *medium* value of HR also have a *high* value of PTT. (T=1).
- *very few* observations with *medium* value of HR also have a *very high* value of PTT. (T=1).

From the global summaries it is possible to observe that half of the observations of medium heart rate have a very small value of PTT. The remaining observations are distributed amongst a few observations that have a small value of PTT and very few observations with medium, high and very high values of PTT. The differences summaries show that for observations with medium heart rate and very small values of PTT, there are more observations for alive patients (fuzzy predicate most) than observations of deceased patients (fuzzy predicate half). For the case of observations with medium heart rate and medium value of PTT, there are more observations of deceased patients (fuzzy predicate few) than alive patients (fuzzy predicate very few).

C. Temporal and Categorical Summaries

Although the previous summaries provide an insight into patients observations, it is be interesting to characterize the heart rate of patients over time. In this section we provide linguistic summaries of patients over time for observations of heart rate, using protoforms (19). We also use protoforms (22) to differentiate patients with different category labels. These summaries are more complete and provide a more complete description of this data. Since this linguistic summary consists of 37 summaries, 12 of which are differences (22), we only present some examples of the obtained summaries. The fuzzy predicates used for the summarizer heart rate are shown in Fig. 2. The linguistic quantifiers and temporal quantifiers are shown in Fig. 3 and Fig. 5.

Differences:

- *few* alive patients have a *low* value of HR *half of the* time, while deceased patients do not. ($d(LS_1, LS_2)$=0.70, T=0.70).
- *half of the* alive patients have a *medium* value of HR, *very few* times, while deceased patients do not. ($d(LS_1, LS_2)$=1T=1).
- *very few* deceased patients have a *very low* value of HR, *half of the* time, while alive patients do not. ($d(LS_1, LS_2)$=0.70, T=1).
- *almost all* deceased patients have a *very high* value of HR, *very few* times, while alive patients do not. ($d(LS_1, LS_2)$=0.70 T=0.70).

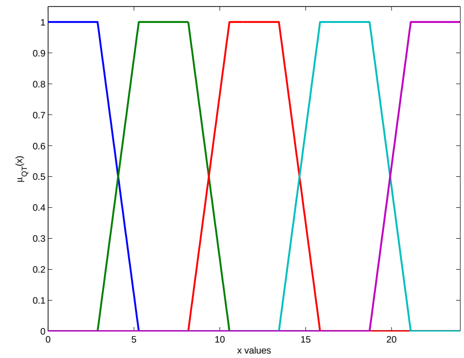

Fig. 5. Membership function for the temporal quantifier μ_{QT}. Left to right '*very few*','*few*','*half*','*most*','*almost all*'.

Global:

- *very few* patients have a *very low* value of HR *half of the* time. (T= 1).
- *few* patients have a *medium* value of HR *most* times. (T= 1)

As expected from the summaries presented in Section V-B, the obtained linguistic summaries of the differences are for value of low, medium and high values of heart rate. In these summaries, it is possible to also have a temporal contextualization of the events. In 6 of the summaries, the event happened very few times (*e.g.* medium heart rate), while 4 of them regard events that happened half of the time (*e.g.* low values of heart rate). In terms of patients 2 of them regard almost all patients, while 8 were about small numbers of patients (very few and few).

VI. Conclusions and Future Work

In this work we provide a simple approach to obtain descriptive linguistic summaries of medical data. We propose an extension of the linguistic summaries protoforms to include categorical data and from these summaries clearly indicate differences exhibited in patients with different class labels. We propose to summarize data using a novel differential form, based on a numerical criterion to compare linguistic summaries. The data set under study is composed of observations on multiple phenomena observed over long time periods for the same patients. To clearly quantify attributes and time, we propose linguistic summaries that provide temporal contextualization. Examples of these new approaches are provided for patients suffering from abdominal septic shock.

In this work we focused on assessing the quality of the linguistic summaries using the truth quality measure. There are several other quality measures [19], [7], future work will transpose them to the considered summaries, so as to further increase their human interpretation and reduce their length.

References

[1] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, no. 1, pp. 69–86, 1982.

[2] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General Systems*, vol. 30, no. 2, pp. 133–154, 2001.

[3] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 813–834, 2000.

[4] L. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Computers & Mathematics with Applications*, vol. 9, no. 1, pp. 149–184, 1983.

[5] B. Bouchon-Meunier and G. Moyse, "Fuzzy linguistic summaries: Where are we, where can we go?" in *Computational Intelligence for Financial Engineering Economics (CIFEr), 2012 IEEE Conference on*, March 2012, pp. 1–8.

[6] A. Niewiadomski, "A type-2 fuzzy approach to linguistic summarization of data," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 1, pp. 198–212, feb. 2008.

[7] D. Wu and J. Mendel, "Linguistic summarization using if–then rules and interval type-2 fuzzy sets," *Fuzzy Systems, IEEE Transactions on*, vol. 19, no. 1, pp. 136–151, feb. 2011.

[8] J. Kacprzyk, A. Wilbik, and S. Zadrożny, "Linguistic summarization of time series using a fuzzy quantifier driven aggregation," *Fuzzy Sets Systems*, vol. 159, no. 12, pp. 1485–1499, June 2008.

[9] J. Kacprzyk and A. Wilbik, "Towards an efficient generation of linguistic summaries of time series using a degree of focus," in *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, june 2009, pp. 1–6.

[10] A. Wilbik, J. Keller, and G. Alexander, "Linguistic summarization of sensor data for eldercare," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, oct. 2011, pp. 2595–2599.

[11] M. Ros, M. Pegalajar, M. Delgado, A. Vila, D. Anderson, J. Keller, and M. Popescu, "Linguistic summarization of long-term trends for understanding change in human behavior," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, june 2011, pp. 2080–2087.

[12] J. Kacprzyk and A. Wilbik, "A comprehensive comparison of time series described by linguistic summaries and its application to the comparison of performance of a mutual fund and its benchmark," in *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, july 2010, pp. 1–8.

[13] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud, "Linguistic summarization of video for fall detection using voxel person and fuzzy logic," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 80–89, 2009.

[14] R. Castillo-Ortega, N. Marin, and D. Sanchez, "Linguistic local change comparison of time series," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, june 2011, pp. 2909–2915.

[15] A. Wilbik and J. M. Keller, "A distance metric for a space of linguistic summaries," *Fuzzy Sets and Systems*, vol. 208, no. 0, pp. 79–94, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165011412001364

[16] A. Wilbik and J. Keller, "A fuzzy measure similarity between sets of linguistic summaries," *Fuzzy Systems, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2012.

[17] A. Wilbik, J. Keller, and J. Bezdek, "Generation of prototypes from sets of linguistic summaries," in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, june 2012, pp. 1 –8.

[18] A. Wilbik, J. M. Keller, and G. L. Alexander, "Similarity evaluation of sets of linguistic summaries," *International Journal of Intelligent Systems*, vol. 27, no. 10, pp. 926–938, 2012. [Online]. Available: http://dx.doi.org/10.1002/int.21555

[19] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," *Information Sciences*, vol. 173, no. 4, pp. 281–304, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S002002550500085X

[20] E. Hanisch, R. Brause, B. Arlt, J. Paetz, and K. Holzer, "The MEDAN database," 2003, http://www.medan.de.

[21] S. Vieira, L. Mendonca, G. Farinha, and J. Sousa, "Metaheuristics for feature selection: Application to sepsis outcome prediction," in *Evolutionary Computation (CEC), 2012 IEEE Congress on*, june 2012, pp. 1–8.

[22] R. D. M. A. Pereira, A. S. Fialho, F. Cismondi, S. M. Vieira, J. a. M. C. Sousa, R. J. Almeida, U. Kaymak, S. R. Reti, M. D. Howell, and S. N. Finkelstein, "Predicting septic shock outcomes in a database with missing data using fuzzy modeling: Influence of pre-processing techniques on real-world data-based classification," in *FUZZ–IEEE*, 2011, pp. 2507–2512.

[23] A. S. Fialho, F. Cismondi, S. M. Vieira, J. a. M. C. Sousa, R. J. Almeida, U. Kaymak, 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 *IPMU*, 2010, pp. 65–74.