Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)

Mining time series data via linguistic summaries of trends by using a modified Sugeno integral based aggregation

Janusz Kacprzyk, *Fellow, IEEE,* Anna Wilbik and Sławomir Zadrożny

***Abstract*— Linguistic summaries as descriptions of trends in time series data are proposed. We further extend our (cf. Kacprzyk, Wilbik and Zadrożny [1], [2], [3], [4], [5]) previous works in which we put forward a new approach to the linguistic summarization of time series. In this paper we basically propose a modification of our previous work on the use of the Sugeno integral developed in [5] by employing a modified fuzzy measure and its related modified Sugeno integral. This gives better results in particular in the case of some more sophisticated and extended types of summaries.**

I. Introduction

Time series are a type of data that is omnipresent and plays a key role in many applications, in virtually all areas of science, economics and technology. A crucial importance of this type of data and an acute need to find effective and efficient methods for its handling has triggered much research and a considerable progress has been made. Notably, statistical methods have played in this respect a considerable role, and recently other methods have been proposed exemplified by those based on neural networks, some biologically inspired paradigms, based on cognitive analysis, etc.

Unfortunately, most of those traditional and new approaches may be viewed to be not human consistent enough in the sense that they do not intend to bridge the essential gap between the human being and the computer. Namely, for the human being the only fully natural means of articulation of some assessments, intentions, etc. is natural language which is strange to the "machine" as algorithm and techniques that are traditionally employed handle numbers and present results in numbers.

This paper is a further step in a new direction to the analysis of time series data that has been proposed in a series of our previous papers (Kacprzyk, Wilbik and Zadrożny [1], [3], [2], [4], [5]). In these papers a new approach to the capturing of the very essence of time series data has been proposed whose essence is the use of natural language descriptions (statements) to describe in a human consistent way how trends in time series data evolve over time, how long some types of behavior last, how rapid changes are, etc.

Basically, in all those works we use the idea of Yager's linguistic data summaries proposed in Yager [6], and then advanced in Kacprzyk and Yager [7], and Kacprzyk, Yager and Zadrożny [8]. In its basic version this approach uses Zadeh's [9] fuzzy logic based calculus of linguistically quantified propositions.

In our context of time series those Yager type summaries can be exemplified by "most trends are short", "most of long trends are slowly increasing" , etc. To derive such linguistic summaries of trends that can help capture what really happens in the time series data under consideration, we proposed first to use Zadeh's classic calculus of linguistically quantified propositions (cf. Kacprzyk, Wilbik and Zadrożny [1]). As a further step, first, new types of linguistic summaries of trends were proposed in Kacprzyk, Wilbik and Zadrożny [2]. Finally, the use of the Sugeno integral was proposed in Kacprzyk, Wilbik and Zadrożny [5].

Some practical calculations using those methods proposed have indicated that some difficulties can occur, notably while using the Sugeno integral for some more sophisticated extended types of linguistic summaries. This was related to problems with the monotonicity of the fuzzy measure employed in the Sugeno integral.

In this paper, while following our previous approach in which the Sugeno integtral has been employed (cf. Kacprzyk, Wilbik and Zadrożny [5]), we put forward its modification by employing a modified fuzzy measure and its related modified Sugeno integral. This gives better results in particular in the case of some more sophisticated and extended types of summaries. we propose some modification of the definition of a fuzzy measure for those extended types of linguistic summaries of trends. This modification implies a new version of the formula for the calculation of the Sugeno integral.

The paper is organized as follows. First we describe the way the trends are extracted from time series and characterized using a set of attributes. Then we briefly remind the basic idea of the original Yager's approach to linguistic data summarization and discuss how it may be used to summarize trends. In the next section we show how these summaries might be interpreted using the concept of a newly defined fuzzy measure and the Sugeno integral. Finally we present some examples of linguistic summaries of a data set.

II. Temporal data and trend analysis

We deal with numerical data that vary over time, and a time series is a sequence of data measured at uniformly spaced

Janusz Kacprzyk and Sławomir Zadrożny are with Systems Research Institute, Polish Academy of Sciences ul. Newelska 6, 01-447 Warsaw, Poland and Warsaw School of Information Technology (WIT) ul. Newelska 6, 01-447 Warsaw, Poland Email: {kacprzyk,zadrozny}@ibspan.waw.pl

Anna Wilbik is with Systems Research Institute, Polish Academy of Sciences ul. Newelska 6, 01-447 Warsaw, Poland Email: wilbik@ibspan.waw.pl

time moments. We will identify trends as linearly increasing, stable or decreasing functions, and therefore represent given time series data as piecewise linear functions. Evidently, the intensity of an increase and decrease (slope) will matter, too. These are clearly partial trends as a global trend in a time series concerns the entire time span of the time series, and there also may be trends that concern parts of the entire time span, but more than a particular window taken into account while extracting partial trends by using the Sklansky and Gonzalez algorithm.

In particular, we use the concept of a uniform partially linear approximation of a time series. Function f is a uniform ε-approximation of a time series or a set of points $\{(x_i, y_i)\}_{i=1,\ldots,n}$ if for a given, context dependent $\varepsilon > 0$, there holds

$$\forall i: \ |f(x_i) - y_i| \leq \varepsilon \qquad (1)$$

We use a modification of the well known Sklansky and Gonzalez [10] effective and efficient algorithm that finds a linear uniform ε-approximation for subsets of points of a time series. The algorithm constructs the intersection of cones starting from point p_i of the time series and including the circle of radius ε around the subsequent data points p_{i+j}, $j = 1, 2, \ldots$, until the intersection of all cones starting at p_i is empty. If for p_{i+k} the intersection is empty, then we construct a new cone starting at p_{i+k-1}. Figures 1 and 2 present the idea of the algorithm. The family of possible solutions is indicated as a gray area. Clearly other algorithms can also be used, and there is a lot of them in the literature.

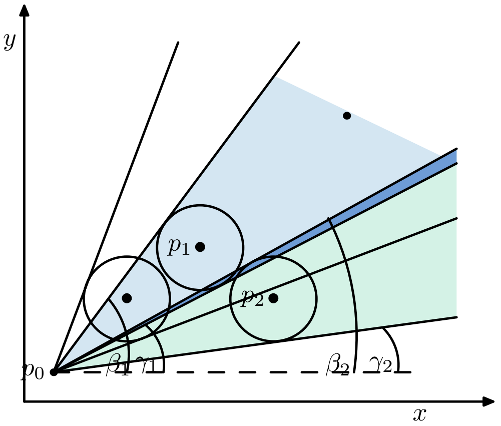

Fig. 1. An illustration of the algorithm for the uniform ε-approximation – the intersection of the cones is indicated by the dark grey area

To present details of the algorithm, let us first introduce the following notation:

- p_0 – a point starting the current cone,
- p_1 – the last point checked in the current cone,
- p_2 – the next point to be checked,
- Alpha_01 – a pair of angles (γ_1, β_1), meant as an interval, defining the current cone as shown in Fig. 1,
- Alpha_02 – a pair of angles defining the cone starting at the point p_0 and inscribing the circle of radius ε around the point p_2 (cf. (γ_2, β_2) in Fig. 1),
- function read_point() reads a next point of data series,
- function find() finds a pair of angles of the cone starting at the point p_0 and inscribing the circle of radius ε around the point p_2.

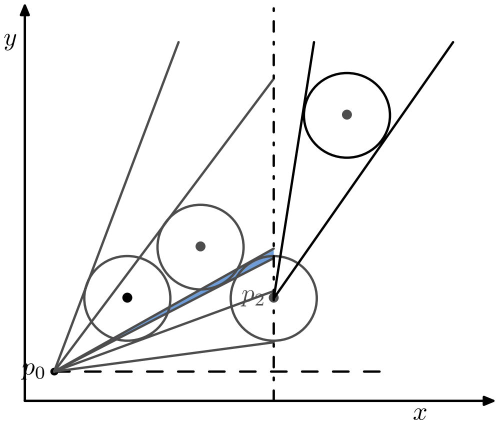

Fig. 2. An illustration of the algorithm for the uniform ε-approximation – a new cone starts in point p_2

The pseudocode of the procedure that extracts the trends is depicted in Fig. 3.

```
read_point(p_0);
read_point(p_1);
while(1)
{
  p_2=p_1;
  Alpha_02=find();
  Alpha_01=Alpha_02;
  do
  {
    Alpha_01 = Alpha_01 $\cap$ Alpha_02;
    p_1=p_2;
    read_point(p_2);
    Alpha_02=find();
  } while(Alpha_01 $\cap$ Alpha_02 $\neq \emptyset$);
  save_found_trend();
  p_0=p_1;
  p_1=p_2;
}
```

Fig. 3. Pseudocode of the modified Sklansky and Gonzalez [10] procedure for extracting trends

The bounding values of Alpha_02 (γ_2, β_2), computed by function find() correspond to the slopes of two lines such that:

- are tangent to the circle of radius ε around point $p_2 = (x_2, y_2)$
- start at the point $p_0 = (x_0, y_0)$

Thus

$$\gamma_2 = arctg\left(\frac{\Delta x \cdot \Delta y - \varepsilon\sqrt{(\Delta x)^2 + (\Delta y)^2 - \varepsilon^2}}{(\Delta x)^2 - \varepsilon^2}\right)$$

and

$$\beta_2 = arctg\left(\frac{\Delta x \cdot \Delta y + \varepsilon\sqrt{(\Delta x)^2 + (\Delta y)^2 - \varepsilon^2}}{(\Delta x)^2 - \varepsilon^2}\right)$$

where $\Delta x = x_0 - x_2$ and $\Delta y = y_0 - y_2$.

The resulting ε-approximation of a group of points `p_0, ..., p_1` is either a single segment, chosen as, e.g. a bisector, or one that minimizes the distance (e.g. assumed as sum of squared errors, SSE) from the approximated points, or the whole family of possible solutions, i.e., the rays of the cone.

This method is effective and efficient as it requires only a single pass through the data. Now we will identify (partial) *trends* with the line segments of the constructed piecewise linear function.

III. Dynamic characteristics of trends

In our approach, while summarizing trends in time series data, we consider the following three aspects:

- dynamics of change,
- duration, and
- variability,

and it should be noted that by trends we mean here global trends, concerning the entire time series (or some, probably large, part of it), not partial trends concerning a small time span (window) taken into account in the (partial) trend extraction phase via the Sklansky and Gonzales [10] algorithm.

In what follows we will briefly discuss these factors.

A. Dynamics of change

Under the term *dynamics of change* we understand the speed of changes. It can be described by the slope of a line representing the trend, (cf. any angle η from the interval $\langle\gamma,\beta\rangle$ in Fig. 1). Thus, to quantify dynamics of change we may use the interval of possible angles $\eta \in \langle -90; 90\rangle$ or their trigonometrical transformation.

However it might be impractical to use such a scale directly while describing trends. Therefore we may use a fuzzy granulation in order to meet the users' needs and task specificity. The user may construct a scale of linguistic terms corresponding to various directions of a trend line as, e.g.:

- quickly decreasing,
- decreasing,
- slowly decreasing,
- constant,
- slowly increasing,
- increasing,
- quickly increasing

Figure 4 illustrates the lines corresponding to the particular linguistic terms.

In fact, each term represents a fuzzy granule of directions. In Batyrshin et al. [11], [12] there are presented many methods of constructing such a fuzzy granulation. The user may define a membership functions of particular linguistic terms depending on his or her needs.

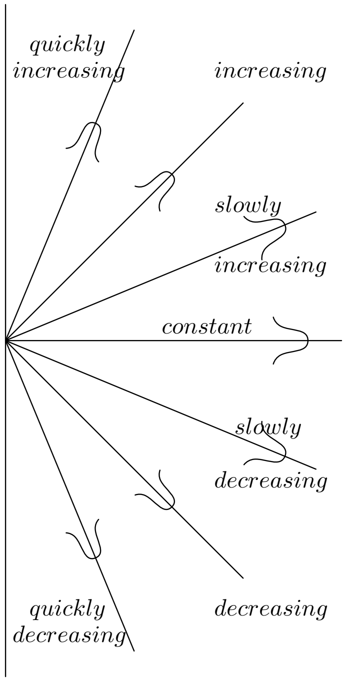

Fig. 4. A visual representation of angle granules defining the dynamics of change

We map a single value α (or the whole interval of angles corresponding to the gray area in Fig. 2) characterizing the dynamics of change of a trend identified using the algorithm shown as a pseudocode in Fig. 3 into a fuzzy set (linguistic label) best matching a given angle. We can use, for instance, some measure of a distance or similarity, cf. the book by Cross and Sudkamp [13]. Then we say that a given trend is, e.g., "decreasing to a degree 0.8", if $\mu_{decreasing}(\alpha) = 0.8$, where $\mu_{decreasing}$ is the membership function of a fuzzy set representing "decreasing" that is a best match for angle α.

B. Duration

Duration describes the length of a single trend, meant as a linguistic variable and exemplified by a "long trend" defined as a fuzzy set whose membership function might be as in Fig. 5 where OX is the time axis divided into appropriate units.

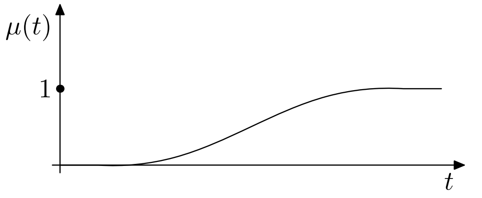

Fig. 5. Example of membership function describing the term "long" concerning the trend duration

The definitions of linguistic terms describing the duration depend clearly on the perspective or purpose assumed by the user.

C. Variability

Variability refers to how "spread out" ("vertically", in the sense of values taken on) a group of data is. The following five statistical measures of variability are widely used in traditional analyses:

- The range (maximum - minimum). Although the range is computationally the easiest measure of variability, it is not widely used, as it is based on only two data points that are extreme. This make it very vulnerable to outliers and therefore may not adequately describe the true variability.
- The interquartile range (IQR) calculated as the third quartile (the third quartile is the 75th percentile) minus the first quartile (the first quartile is the 25th percentile) that may be interpreted as representing the middle 50% of the data. It is resistant to outliers and is computationally as easy as the range.
- The variance is calculated as $\frac{\sum_i (x_i - \bar{x})^2}{n}$, where $\bar{x}$ is the mean value.
- The standard deviation – a square root of the variance. Both the variance and the standard deviation are affected by extreme values.
- The mean absolute deviation (MAD), calculated as $\frac{\sum_i |x_i - \bar{x}|}{n}$. It is not frequently encountered in mathematical statistics. This is essentially because while the mean deviation has a natural intuitive definition as the "mean deviation from the mean" but the introduction of the absolute value makes analytical calculations using this statistic much more complicated.

We propose to measure the variability of a trend as the distance of the data points covered by this trend from a linear uniform ε-approximation (cf. Section II) that represents a given trend. For this purpose we propose to employ a distance between a point and a family of possible solutions, indicated as a gray cone in Fig. 1. Equation (1) assures that the distance is definitely smaller than ε. We may use this information for the normalization. The normalized distance equals 0 if the point lays in the gray area. In the opposite case it is equal to the distance to the nearest point belonging to the cone, divided by ε. Alternatively, we may bisect the cone and then compute the distance between the point and this ray.

Similarly as in the case of dynamics of change, we find for a given value of variability obtained as above a best matching fuzzy set (linguistic label) using, e.g., some measure of a distance or similarity, cf. the book by Cross and Sudkamp [13]. Again the measure of variability is treated as a linguistic variable and expressed using linguistic terms (labels) modeled by fuzzy sets defined by the user.

IV. Linguistic data summaries

A linguistic summary is meant as a (usually short) natural language like sentence (or some sentences) that subsumes the very essence of a set of data (cf. Kacprzyk and Zadrożny [14], [15]). This data set is numeric and usually large, not comprehensible in its original form by the human being. In Yager's approach (cf. Yager [6], Kacprzyk and Yager [7], and Kacprzyk, Yager and Zadrożny [8]) the following perspective for linguistic data summaries is assumed:

- $Y = \{y_1, \ldots, y_n\}$ is a set of objects (records) in a database, e.g., the set of workers;
- $A = \{A_1, \ldots, A_m\}$ is a set of attributes characterizing objects from Y, e.g., salary, age, etc. in a database of workers, and $A_j(y_i)$ denotes a value of attribute A_j for object y_i.

A linguistic summary of a data set consists of:

- a summarizer P, i.e. an attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_j (e.g. "low salary" for attribute "salary");
- a quantity in agreement Q, i.e. a linguistic quantifier (e.g. most);
- truth (validity) $\mathcal{T}$ of the summary, i.e. a number from the interval $[0, 1]$ assessing the truth (validity) of the summary (e.g. 0.7); usually, only summaries with a high value of $\mathcal{T}$ are interesting;
- optionally, a qualifier R, i.e. 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").

Thus, a linguistic summary may be exemplified by

$$\mathcal{T}(\text{most of employees earn low salary}) = 0.7 \tag{2}$$

or, in a richer (extended) form, including a qualifier (e.g. young), by

$$\mathcal{T}(\text{most of young employees earn low salary}) = 0.9 \tag{3}$$

Thus, basically, the core of a linguistic summary is a *linguistically quantified proposition* in the sense of Zadeh [9] which, for (2), may be written as

$$Qy\text{'s are } P \tag{4}$$

and for (3), may be written as

$$QRy\text{'s are } P \tag{5}$$

Then, $\mathcal{T}$, i.e., the truth (validity) of a linguistic summary, directly corresponds to the truth value of (4) or (5). This may be calculated by using either original Zadeh's calculus of linguistically quantified propositions (cf. [9]), or other interpretations of linguistic quantifiers.

The truth values (from $[0, 1]$) of (4) and (5) are calculated, respectively, as

$$T(Qy\text{'s are } P) = \mu_Q \left(\frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right) \tag{6}$$

$$T(QRy\text{'s are } P) = \mu_Q \left(\frac{\sum_{i=1}^{n} (\mu_R(y_i) \wedge \mu_P(y_i))}{\sum_{i=1}^{n} \mu_R(y_i)} \right) \tag{7}$$

where Q is a fuzzy set representing the linguistic quantifier in the sense of Zadeh [9].

V. Protoforms of linguistic trend summaries

It was shown by Kacprzyk and Zadrożny [14] that Zadeh's [16] concept of the protoform is convenient for dealing with linguistic summaries. This approach is also employed here.

Basically, a protoform is defined as a more or less abstract prototype (template) of a linguistically quantified proposition. Then, the summaries mentioned above might be represented by two types of the protoforms:

- Summaries based on frequency:
  - a protoform of a short form of linguistic summaries:

    $$Q \text{ trends are } P \tag{8}$$

    and exemplified by:

    Most of trends have *a large variability*

  - a protoform of an extended form of linguistic summaries:

    $$QR \text{ trends are } P \tag{9}$$

    and exemplified by:

    Most of *slowly decreasing trends* have *a large variability*

- Duration based summaries:
  - a protoform of a short form of linguistic summaries:

    $$\text{The trends that took } Q \text{ time are } P \tag{10}$$

    and exemplified by:

    The trends that took *most* time have *a large variability*

  - a protoform of an extended form of linguistic summaries:

    $$R \text{ trends that took } Q \text{ time are } P \tag{11}$$

    and exemplified by:

    Slowly decreasing trends that took *most* time have *a large variability*

By the very essence of our approach, we need to perform a linguistic quantifier driven aggregation.

The truth degrees T of the frequency based summaries (8)–(9) can be directly computed using Zadeh's calculus of linguistically quantified propositions, in particular the formulas (6) and (7) are of use. To be more specific, the truth values (from $[0, 1]$) of (8) and (9) are calculated, respectively, as

$$\mathcal{T}(Qy\text{'s are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{12}$$

and

$$\mathcal{T}(QRy\text{'s are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}(\mu_R(y_i)\wedge\mu_P(y_i))}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{13}$$

where, here and later on, we assume, for obvious reasons, that the respective denominators are not equal to zero. Otherwise we need to resort to slightly modlified formulas, but this will not be considered in this paper.

The calculation of the truth values of duration based summaries is more complicated and requires a different approach. In the case of a summary "the trends that took Q time are P" we should calculate the time that is taken by "trend is P". It is obvious then when "trend is P" to degree 1, then we can use the whole time taken by this trend. However, what to do if "trend is P" is to some degree (less than 1)? We propose to only take a part of the time defined by the degree to which "trend is P". In other words, we calculate this time as $\mu(y_i)t_{y_i}$, where t_{y_i} is the duration of trend y_i. The value obtained (the duration of such trends that "trend is P") is then normalized by dividing it by the overall time T. Finally, we may calculate to which degree the time taken by such trends that "trend is P" is Q.

We proceed in a similar way in case of the extended form of linguistic summaries. Thus, we obtain the following formulas:

- for the short form of the duration based summaries (10):

  $$\mathcal{T}(y \text{ that took } Q \text{ time are } P) = \\ = \mu_Q\left(\frac{1}{T}\sum_{i=1}^{n}\mu_P(y_i)t_{y_i}\right) \tag{14}$$

  where T is the total time of the summarized trends and t_{y_i} is the duration of the ith trend;
- for the extended form of the duration based summaries (11):

  $$\mathcal{T}(Ry \text{ that took } Q \text{ time are } P) = \\ = \mu_Q\left(\frac{\sum_{i=1}^{n}(\mu_R(y_i)\wedge\mu_P(y_i))t_{y_i}}{\sum_{i=1}^{n}\mu_R(y_i)t_{y_i}}\right) \tag{15}$$

  where t_{y_i} is the duration of the ith trend.

One can notice that the procedure outlined above implies simple, highly intuitively appealing formulas in the case of the frequency based summaries while more complicated, yet not intuitively appealing formulas in the case of the duration based summaries. We will show in this paper that by using a modication of the Sugeno integral based aggregation uniform, intuitively appealing formulas can be obtained in both cases.

VI. Linguistic summary interpretation via the Sugeno integral

As mentioned in the previous section, the use of Zadeh's calculus of linguistically quantified propositions is well justified in the case of the frequency based summaries while may lead to some questionable results in the case of the duration based summaries. A Sugeno integral based aggregation may help as proposed by Kacprzyk, Wilbik and Zadrożny [5]; this will be outlined below.

Let us start with a brief recall of the basics of the Sugeno integral. Let $X = \{x_1, \ldots, x_n\}$ be a finite set. Then, (cf., e.g., [17]) a *fuzzy measure* on X is a set function $\mu : \mathcal{P}(X) \longrightarrow$ [0,1] such that:

$$\begin{array}{l} \mu(\emptyset) = 0, \mu(X) = 1 \\ \text{if } A \subseteq B \text{ then } \mu(A) \leq \mu(B), \qquad \forall A, B \in \mathcal{P}(X) \end{array} \tag{16}$$

where $\mathcal{P}(X)$ denotes a set of all subsets of X.

Let μ is a fuzzy measure on X. The *discrete Sugeno integral* of a function $f : X \longrightarrow [0, 1],\ f(x_i) = a_i$, with respect to μ is a function $S_\mu : [0, 1]^n \longrightarrow [0, 1]$ such that

$$S_\mu(a_1, \ldots, a_n) = \max_{i=1,\ldots,n} (a_{\sigma(i)} \wedge \mu(B_i)) \tag{17}$$

where $\wedge$ stands for the minimum, σ is such a permutation of $\{1, \ldots, n\}$ that $a_{\sigma(i)}$ is the i-th smallest element from among the a_i's and $B_i = \{x_{\sigma(i)}, \ldots, x_{\sigma(n)}\}$.

We can treat function f as a membership function of a fuzzy set $F \in \mathcal{F}(X)$, where $\mathcal{F}(X)$ denotes a family of fuzzy sets defined in X. Then the Sugeno integral can be equivalently defined as a function $S_\mu : \mathcal{F}(X) \longrightarrow [0, 1]$ such that

$$S_\mu(F) = \max_{\alpha_i \in \{a_1,\ldots,a_n\}} (\alpha_i \wedge \mu(F_{\alpha_i})) \tag{18}$$

where F_{α_i} is the α-cut of F and the meaning of other symbols is as in (17).

The fuzzy measure and the Sugeno integral may be intuitively interpreted in the context of multicriteria decision making (MCDM) where we have a set of criteria and some options (decisions) characterized by the degree of satisfaction of particular criteria. In such a setting X is a set of criteria and μ expresses the importance of each subset of criteria, i.e., how the satisfaction of a given subset of criteria contributes to the overall evaluation of the option. Then the properties of the fuzzy measure (16) properly require that the satisfaction of all criteria makes an option fully satisfactory and that the more criteria are satisfied by an option the better its overall evaluation. Finally the set F represents an option and $\mu_F(x)$ defines the degree to which it satisfies the criterion x. Then the Sugeno integral may be interpreted as an aggregation operator yielding an overall evaluation of option F in terms of its satisfaction of the set of criteria X. In such a context the formula (18) may interpreted as follows:

> find a subset of criteria of the highest possible importance (expressed by μ) such that at the same time minimal satisfaction degree of all these criteria by the option F is as high as possible (expressed by α)
> and take the minimum of these two degrees as the overall evaluation of the option F. (19)

Now we will explain how various linguistic summaries discussed in the previous section may be interpreted using the Sugeno integral. The linguistic quantifier Q is still defined as in Zadeh's calculus as a fuzzy set in [0,1], exemplified by (22). We will assume that Q is a regular monotone and nondecreasing quantifier:

$$\mu(0) = 0, \quad \mu(1) = 1 \tag{20}$$

$$x_1 \leq x_2 \Rightarrow \mu_Q(x_1) \leq \mu_Q(x_2) \tag{21}$$

exemplified by

$$\mu_Q(x) = \begin{cases} 1 & \text{for } x \geq 0.8 \\ 2x - 0.6 & \text{for } 0.3 < x < 0.8 \\ 0 & \text{for } x \leq 0.3 \end{cases} \tag{22}$$

The truth value of particular summaries is calculated using the Sugeno integral (18). For simple types of summaries we are in a position to provide the interpretation similar to this given above for the MCDM. For this purpose we will identify the set of criteria X with a set of trends while an option F will be the whole time series under consideration characterized in terms of how well its trends satisfy P.

Unfortunately, a direct application of the line of reasoning shown above may lead sometimes to problems with the monotonicity of the measure used, in particular in the case of an extended form of a summary. That is why we propose in this paper some modifications to the definition of a fuzzy measure to overcome these possible difficulties.

a) Simple frequency based summaries defined by (8): The truth value of this type of a summary may be expressed as:

$$\mathcal{T}(Q \text{ trends are } P) = \max_{\alpha \in \{a_1,\ldots,a_n\}} \left(\alpha \wedge \mu_Q \left(\frac{|P_\alpha|}{|X\ |} \right) \right) \tag{23}$$

Thus, referring to (19), the truth value is determined by looking for a subset of trends of high enough a cardinality as required by the semantics of the quantifier Q and such that all these trends "are P" to the highest possible degree.

b) Extended frequency based summaries defined by (9): For the simple form of summaries the formula is not complicated. However it is more difficult to find a proper fuzzy measure for an extended type of summaries, where we need to limit our universe of discourse to trends that "are R".

The truth value of this type of a summary may be expressed as:

$$\mathcal{T}(QR \text{ trends are } P) = \\ = \max_{\beta \in \{b_1,\ldots,b_k\}} (\max_{\alpha \in \{a_1,\ldots,a_n\}} (\alpha \wedge \ \mu_Q \left(\frac{|\ (P \cap R_\beta)_\alpha\ |}{|\ R_\beta\ |} \right))) \tag{24}$$

First we have to specify our "universe (of discourse)", a subset of such trends that all "are R". Then, referring to (19), for each such a "universe" we calculate the truth value, determined by looking for a subset of trends of high enough a cardinality as required by the semantics of quantifier Q in comparison to the ”universe” and such that all these trends "are P and R" to the highest possible degree. As the final result we take the maximum value of all truth values obtained for each "universe".

c) Simple duration based summaries defined with (10): The truth value of this type of a summary may be expressed as:

$$\mathcal{T}(\text{Trends that took } Q \text{ time are } P) = \\ = \max_{\alpha \in \{a_1,\ldots,a_n\}} \left(\alpha \wedge \mu_Q \left(\frac{\sum_{i: x_i \in P_\alpha} \text{time}(x_i)}{\sum_{i: x_i \in X} \text{time}(x_i)} \right) \right) \tag{25}$$

Thus, referring to (19) the truth value is determined by looking for a subset of trends such that their total duration with respect to the duration of the whole time series is long enough as required by the semantics of the quantifier Q and such that all these trends "are P" to the highest possible degree.

d) Extended duration based summaries defined with (11): The truth value of this type of a summary may be expressed as:

$$\begin{aligned}
&\mathcal{T}(R \text{ trends that took } Q \text{ time are } P) = \\
&= \max_{\beta\in\{b_1,\ldots,b_k\}} \left(\max_{\alpha\in\{a_1,\ldots,a_n\}} (\alpha \wedge \right. \\
&\left. \wedge\, \mu_Q \left(\frac{\sum_{i:x_i\in(P\cap R_\beta)_\alpha} \text{time}(x_i)}{\sum_{i:x_i\in R_\beta} \text{time}(x_i)} \right) \right) \Bigg) \qquad (26)
\end{aligned}$$

Due to the properties (20)–(21) of the quantifiers employed it is obvious that all μ's defined above for particular types of summaries satisfy the axioms (16) of the fuzzy measure.

VII. Example

Let us assume that from some given data we have extracted trends listed in Table I, e.g. using the algorithm shown in Fig. 3. We assume the granulation of dynamics of change presented in Section III-A.

TABLE I
Trends extracted

id	dynamics of change (α in degrees)	duration (time units)	variability ([0,1])
1	25	15	0.2
2	-45	1	0.3
3	75	2	0.8
4	-40	1	0.1
5	-55	1	0.7
6	50	2	0.3
7	-52	1	0.5
8	-37	2	0.9
9	15	5	0.0

We can consider the following simple frequency based summary:

$$Most \text{ of trends are } decreasing \qquad (27)$$

In this summary *most* is the linguistic quantifier Q. The membership function is as in (22).

"Trends are decreasing" is a summarizer P with the membership function of the "decreasing" term given as in (28). Let us recall, that for brevity we identify summarizers and qualifiers with the linguistic terms they contain.

$$\mu_P(\alpha) = \begin{cases} 0 & \text{for } \alpha \le -65 \\ 0.066\alpha + 4.333 & \text{for } -65 < \alpha < -50 \\ 1 & \text{for } -50 \le \alpha \le -40 \\ -0.05\alpha - 1 & \text{for } -40 < \alpha < -20 \\ 0 & \text{for } \alpha \ge -20 \end{cases} \qquad (28)$$

n is the number of all trends, i.e., in this example $n = |X|$=9.

The truth value of (27) is computed according to (18) and (23) that yields:

$$\begin{aligned}
&\mathcal{T}(Most \text{ of the trends are } decreasing) = \\
&= \max_{\alpha_i\in\{a_1,\ldots,a_n\}} \left(\alpha_i \wedge \mu_Q \left(\frac{|P_\alpha|}{|X\ |} \right) \right) = 0.511
\end{aligned}$$

If the truth value is computed according to the (12) we obtain:

$$\begin{aligned}
&\mathcal{T}(Most \text{ of the trends are } decreasing) = \\
&= \mu_Q \left(\frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right) = 0.601
\end{aligned}$$

If we assume the extended form, we may have the following summary:

$$Most \text{ of } short \text{ trends are } decreasing \qquad (29)$$

Again, *most* is the linguistic quantifier Q with its membership function given as (22). *"Trends are decreasing"* is a summarizer P as in the previous example. *"Trend is short"* is the qualifier R. We define the membership function $\mu_R(t)$ as follows:

$$\mu_R(t) = \begin{cases} 1 & \text{for } t \le 1 \\ -\frac{1}{2}t + \frac{3}{2} & \text{for } 1 < t < 3 \\ 0 & \text{for } t \ge 3 \end{cases} \qquad (30)$$

The truth value of (29) is computed using the formula (18) and (24):

$$\begin{aligned}
&\mathcal{T}(Most \text{ of } short \text{ trends are } decreasing) = \\
&= \max_{\beta\in\{b_1,\ldots,b_k\}} \left(\max_{\alpha\in\{a_1,\ldots,a_n\}} \left(\alpha \wedge \mu_Q \left(\frac{|(P\cap R_\beta)_\alpha|}{|R_\beta|} \right) \right) \right) \\
&= 0.54
\end{aligned}$$

If the truth value is computed according to the (13) we obtain:

$$\begin{aligned}
&\mathcal{T}(Most \text{ of } short \text{ trends are } decreasing) = \\
&= \mu_Q \left(\frac{\sum_{i=1}^{n} (\mu_R(y_i) \wedge \mu_P(y_i))}{\sum_{i=1}^{n} \mu_R(y_i)} \right) = 0,822
\end{aligned}$$

On the other hand, we may have the following simple duration based linguistic summary:

$$\text{Trends that took } most \text{ time are } slowly\ increasing \qquad (31)$$

"Trends are slowly increasing" is the summarizer P with the membership function $\mu_P(\alpha)$ defined as follows:

$$\mu_P(\alpha) = \begin{cases} 0 & \text{for } \alpha \le 5 \\ 0.1\alpha - 0.5 & \text{for } 5 < \alpha < 15 \\ 1 & \text{for } 15 \le \alpha \le 20 \\ -0.05\alpha + 2 & \text{for } 20 < \alpha < 40 \\ 0 & \text{for } \alpha \ge 40 \end{cases} \qquad (32)$$

The linguistic quantifier *most* is defined as previously. The truth value of (31) is computed via the formula (18) and (25) and we obtain:

$$\begin{aligned}
&\mathcal{T}(\text{Trends that took } most \text{ time are } slowly\ increasing) = \\
&= \max_{\alpha_i\in\{a_1,\ldots,a_n\}} \left(\alpha_i \wedge \mu_Q \left(\frac{\sum_{x_i\in P_\alpha} \text{time}(x_i)}{\sum_{i:x_i\in X} \text{time}(x_i)} \right) \right) \\
&= 0.733
\end{aligned}$$

If the truth value is computed according to the (14) we obtain:

$$\begin{aligned}&\mathcal{T}(\text{Trends that took } most \text{ time are } slowly\ increasing) = \\ &= \quad \mu_Q\left(\frac{1}{T}\sum_{i=1}^{n}\mu_P(y_i)t_{y_i}\right) = 0,733\end{aligned}$$

Finally, we may consider an extended form of duration based summaries, here exemplified by:

$$\begin{aligned}&\text{Trends } with\ a\ low\ variability \text{ that took } most \text{ of} \\ &\qquad\qquad \text{the time are } slowly\ increasing \qquad (33)\end{aligned}$$

Again, *most* is the linguistic quantifier and *"trends are slowly increasing"* is summarizer P, with a membership function defined as in the previous example. *"Trends have a low variability"* is the qualifier R. The membership function $\mu_R(v)$ is given as follows:

$$\mu_R(v) = \begin{cases} 1 & \text{for } v \le 0.2 \\ -5v + 2 & \text{for } 0.2 < v < 0.4 \\ 0 & \text{for } v \ge 0.4 \end{cases} \qquad (34)$$

The truth value of (33) is computed according to the formula (18) and (26) and we obtain:

$$\begin{aligned}&\mathcal{T}(\text{Trends } with\ low\ variability \text{ that took } most \text{ of} \\ &\qquad \text{the time are } slowly\ increasing) = \\ &= \quad \max_{\beta\in\{b_1,\ldots,b_k\}}\left(\max_{\alpha\in\{a_1,\ldots,a_n\}}\left(\alpha \wedge\right.\right. \\ &\qquad \left.\left.\wedge\, \mu_Q\left(\frac{\sum_{i:x_i\in(P\cap R_\beta)_\alpha}\text{time}(x_i)}{\sum_{i:x_i\in R_\beta}\text{time}(x_i)}\right)\right)\right) \\ &= \quad 0.75\end{aligned}$$

If the truth value is computed according to the (15) we obtain:

$$\begin{aligned}&\mathcal{T}(\text{Trends } with\ low\ variability \text{ that took } most \text{ of} \\ &\qquad \text{the time are } slowly\ increasing) = \\ &= \quad \mu_Q\left(\frac{\sum_{i=1}^{n}(\mu_R(y_i)\wedge\mu_P(y_i))t_{y_i}}{\sum_{i=1}^{n}\mu_R(y_i)t_{y_i}}\right) = 1\end{aligned}$$

VIII. Concluding remarks

We have proposed a further extension of our previous works (cf. Kacprzyk, Wilbik and Zadrożny [1], [2], [3], [4], [5]) in which we proposed a new approach to the linguistic summarization of time series. In this paper we have taken as a point of departure our recent work in this direction [5] in which a Sugeno integral based aggregation has been employed, and have extended it by employing a modified fuzzy measure and its related modified Sugeno integral. This gives better results in particular in the case of some more sophisticated and extended types of summaries.

References

[1] J. Kacprzyk, A. Wilbik, and S. Zadrożny, "Linguistic summarization of trends: A fuzzy logic based approach," in *Proceedings of the 11th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems*, 2006, pp. 2166–2172, Paris, France, July 2-7, 2006.

[2] ——, "On some types of linguistic summaries of time series," in *Proceedings of the 3rd International IEEE Conference"Intelligent Systems"*. IEEE Press, 2006, pp. 373–378, London, UK, September 4-6, 2006.

[3] ——, "A linguistic quantifier based aggregation for a human consistent summarization of time series," in *Soft Methods for Integrated Uncertainty Modelling*, J. Lawry, E. Miranda, A. Bugarin, S. Li, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, Eds. Springer-Verlag, Berlin and Heidelberg, 2006, pp. 186–190.

[4] ——, "Capturing the essence of a dynamic behavior of sequences of numerical data using elements of a quasi-natural language," in *Proceedings of the "2006 IEEE International Conference on Systems, Man, and Cybernetics"*. IEEE Press, 2006, pp. 3365–3370, Taipei, Taiwan, October 8 – 11, 2006.

[5] ——, "Linguistic summaries of time series via a quantifier based aggregation using the Sugeno integral," in *Proceedings of 2006 IEEE World Congress on Computational Intelligence*. IEEE Press, 2006, pp. 3610–3616, Vancouver, BC, Canada, July 16-21, 2006.

[6] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, pp. 69–86, 1982.

[7] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General Systems*, vol. 30, pp. 33–154, 2001.

[8] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of Applied Mathematics and Computer Science*, vol. 10, pp. 813–834, 2000.

[9] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Computers and Mathematics with Applications*, vol. 9, pp. 149–184, 1983.

[10] J. Sklansky and V. Gonzalez, "Fast polygonal approximation of digitized curves," *Pattern Recognition*, vol. 12, no. 5, pp. 327–331, 1980.

[11] I. Batyrshin, "On granular derivatives and the solution of a granular initial value problem," *International Journal Applied Mathematics and Computer Science*, vol. 12, no. 3, pp. 403–410, 2002.

[12] I. Batyrshin and L. Sheremetov, "Perception based functions in qualitative forecasting," in *Perception-based Data Mining and Decision Making in Economics and Finance*, I. Batyrshin, J. Kacprzyk, L. Sheremetov, and L. Zadeh, Eds. Springer-Verlag, Berlin and Heidelberg, 2006.

[13] V. Cross and T. Sudkamp, *Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications*. Heidelberg and New York: Springer-Verlag, 2002.

[14] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools," *Information Sciences*, vol. 173, pp. 281–304, 2005.

[15] ——, "Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining," in *Do Smart Adaptive Systems Exist?*, B. Gabrys, K. Leiviska, and J. Strackeljan, Eds. Berlin, Heidelberg, New York: Springer, 2005, pp. 321–339.

[16] L. A. Zadeh, "A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform," in *BISC Seminar*, University of California, Berkeley, 2002.

[17] M. Grabisch, "Fuzzy integral as a flexible and interpretable tool of aggregation," in *Aggregation and Fusion of Imperfect Information*, B. Bouchon-Meunier, Ed. Heidelberg, New York: Physica–Verlag, 1998, pp. 51–72, studies in Fuzziness and Soft Computing.