A FUZZY LOGIC APPROACH TO THE LINGUISTIC SUMMARIZATION OF TIME SERIES

Janusz KACPRZYK, Anna WILBIK, Sławomir ZADROŻNY

Systems Research Institute
Polish Academy of Sciences
E-mail: {kacprzyk,wilbik,zadrozny}@ibspan.waw.pl

Abstract

We consider our approach to the linguistic summarization of time series data proposed in our previous papers. We summarize trends identified here with straight segments of a piecewise linear approximation of time series. Then we employ, as a set of features, the duration, dynamics of change and variability, and assume different, human consistent granulations of their values. The problem boils down to a linguistic quantifier driven aggregation of partial trends that is done via the classic Zadeh’s calculus of linguistically quantified propositions. We show an application to linguistic summarization of time series data on daily quotations of an investment fund over an eight year period.

Keywords: linguistic summary, trend analysis, fuzzy logic, computing with words.

1 Introduction

A linguistic data (base) summary is meant as a concise, human-consistent description of a (numerical) data set. This concept has been introduced by Yager [22] and then further developed by Kacprzyk and Yager [14], and Kacprzyk, Yager and Zadrozny [15]. In this approach the contents of a database is summarized via a natural language like expression semantics of which is provided in the framework of Zadeh’s calculus of linguistically quantified propositions [26].
Since data sets in most nontrivial cases are large, if not huge, it is very difficult for a human being to capture and understand their contents. A natural language like description would be very helpful as natural language is the only fully natural means of articulation and communication for a human being. In this paper we consider a specific type of data, namely time series, i.e. a certain real valued function of time. For a manager, stock exchange players, etc., it might be convenient and useful to obtain a brief, natural language like description of trends present in the data on a company performance, stock exchange quotations, etc. over a certain period of time.

Though statistical methods exhibit their strength in such cases, and are often used, in our case we attempt to derive (quasi)natural language like descriptions that should be considered to be an additional form of data description of a remarkably high human consistency because – as we have already indicated – for a human being the only fully natural means of articulation, communication, etc. is natural language. Hence, our approach is not meant to replace the classical statistical analyses but rather serve as an additional form of data description characterized by its high human consistency.

The summaries of time series we propose refer in fact to the summaries of trends identified here with straight line segments of a piece-wise linear approximation of time series. Thus, the first step is the construction of such an approximation. For this purpose we use a modified version of the simple, easy to use Sklansky and Gonzalez algorithm presented in [20].

Then we employ a set of features (attributes) to characterize the trends such as the slope of the line, the fairness of approximation of the original data points by line segments and the length of a period of time comprising the trend.

Basically the summaries proposed by Yager are interpreted in terms of the number or proportion of elements possessing a certain property. In the framework considered here a summary might look like: “Most of the trends are short” or in a more sophisticated form: “Most long trends are increasing”. Such expressions are easily interpreted using Zadeh’s calculus of linguistically quantified propositions. The most important element of this interpretation is a linguistic quantifier exemplified by “most”. In Zadeh’s [26] approach it is interpreted in terms of a proportion of elements possessing a certain property (e.g., a length of a trend) among all the elements considered (e.g., all trends).

In Kacprzyk, Wilbik and Zadrożyński [9] we proposed to use Yager’s linguistic summaries, interpreted in the framework of Zadeh’s calculus of linguistically quantified propositions, for the summarization of time series. In our further papers (cf. Kacprzyk, Wilbik and Zadrożyński [11, 12, 13]) we proposed, first, another type of summaries that does not use the linguistic quantifier based
aggregation over the number of trends but over the time instants they take altogether. For example, such a summary can be: “Trends taking most of the time are increasing” or “Increasing trends taking most of the time are of a low variability”. Such summaries do not directly fit the framework of the original Yager’s approach and to overcome this difficulty we generalize our previous approach by modelling the linguistic quantifier based aggregation both over the number of trends as well over the time they take using first the Sugeno integral and then the Choquet integral. All these approaches have been proposed using a unified perspective given by Kacprzyk and Zadrozny [16] that is based on Zadeh’s [27] protoforms.

In this paper we will basically employ the classic Zadeh’s calculus of linguistically quantified propositions. However, we will analyse here only the frequency based summaries.

The paper is in line with some modern approaches to a human consistent summarization of time series. First of all, one should cite here the works of Batyrshin and his collaborators [1, 2]. Basically, they consider the problem in terms of devising a rule base, and then assume a different approach to linguistic granulation. Chiang, Chow and Wang’s [6] approach, though it basically addresses a problem that is similar in spirit, is also different.

To see our approach in a proper perspective it may be expedient to refer to a interesting project coordinated by the University of Aberdeen, UK, SumTime, an EPSRC Funded Project for Generating Summaries of Time Series Data (cf. www.csd.abdn.ac.uk/research/sumtime/). The essence of this project can be summarized by the citation from its Web site: “Our goal is to develop technology for producing English summary descriptions of a time-series data set. Currently there are many visualisation tools for time-series data, but techniques for producing textual descriptions of time-series data are much less developed. Some systems have been developed in the natural-language generation (NLG) community for tasks such as producing weather reports from weather simulations, or summaries of stock market fluctuations, but such systems have not used advanced time-series analysis techniques. Our goal is to develop better technology for producing summaries of time-series data by integrating leading-edge time-series and NLG technology”.

Basically, the essence of this project is close in intent and spirit to our works. However, the type of summaries they generate is different, not accounting for an inherent imprecision of natural language. A good example is here the case of weather prediction that is one of the main application areas in that project. For instance, cf. Sripada et al. [21], linguistic summaries related to wind direction and speed can be:
WSW (West of South West) at 10-15 knots increasing to 17-22 knots early morning, then gradually easing to 9-14 knots by midnight.

- During this period, spikes simultaneously occur around 00:29, 00:54, 01:08, 01:21, and 02:11 (o’clock) in these channels.

Similar linguistic summaries have been obtained for time series data on blood pressure, gas turbines, etc.

Notice that these linguistic description of time series data concerning wind directions and speed do provide a higher human consistency as natural language is used but they capture imprecision of natural language to a very limited extent. In our approach this will be overcome to a considerable extent.

2 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 time moments. We 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 [20] algorithm.

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

$$ \forall i : |f(x_i) - y_i| \leq \epsilon $$

and, clearly, if $f$ is linear, then such an approximation is a linear uniform $\epsilon$-approximation.

We use a modification of the well known, effective and efficient Sklansky and Gonzalez [20] algorithm that finds a linear uniform $\epsilon$-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 a circle of radius $\epsilon$ 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(a) and 1(b) 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 (e.g. [18, 19]).

(a) the intersection of the cones is indicated by the dark grey area
(b) a new cone starts in point $p_2$

Figure 1. An illustration of the algorithm for the uniform $\varepsilon$-approximation

To present details of the algorithm, let us first denote:

- $p_0$ – a point starting the current cone,
- $p_1$ – the last point checked in the current cone,
- $p_2$ – the next point to be checked,
- $\text{Alpha}_01$ – a pair of angles $(\gamma_1, \beta_1)$, meant as an interval, that defines the current cone as shown in Figure 1(a),
- $\text{Alpha}_02$ – a pair of angles of the cone starting at the point $p_0$ and inscribing the circle of radius $\varepsilon$ around the point $p_2$ (cf. $(\gamma_2, \beta_2)$ in Figure 1(a)),
- function $\text{read\_point}()$ reads a next point of data series,
- function $\text{find}()$ finds a pair of angles of the cone starting at the point $p_0$ and inscribing the circle of radius $\varepsilon$ around the point $p_2$.

A pseudocode of the algorithm that extracts trends is depicted in Figure 2. The bounding values of $\text{Alpha}_02$ $(\gamma_2, \beta_2)$, computed by function $\text{find}()$ correspond to the slopes of two lines that:

- are tangent to the circle of radius $\varepsilon$ around point $p_2 = (x_2, y_2)$
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;
}

Figure 2. Pseudocode of the modified Sklansky and Gonzalez [20] algorithm for extracting trends

- 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 linear $\varepsilon$-approximation of a group of points $p_0, \ldots, p_{-1}$ is either a single segment, chosen as, e.g., a bisector of the cone, or one that minimizes the distance (e.g., the 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. Among some other approaches, works of Keogh and his collaborators [18, 19] should be cited, and some of their ideas will be employed in our further works.

3 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 a 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 [20] algorithm mentioned above.

In what follows we will briefly discuss these factors.

3.1 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 $\eta$ from the interval $(\gamma, \beta)$ in Fig. 1(a)). Thus, to quantify dynamics of change we may use the interval of possible angles $\eta \in (-90^\circ; 90^\circ)$ 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 3 illustrates the lines corresponding to the particular linguistic terms.

![Diagram showing angle granules defining the dynamics of change](image)

**Figure 3.** A visual representation of angle granules defining the dynamics of change

In fact, each term represents a fuzzy granule of directions. In Batyrshin et al. [1, 2] 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.

We map a single value $\alpha$ (or the whole interval of angles corresponding to the gray area in Fig. 1(b)) characterizing the dynamics of change of a trend identified using the algorithm shown as a pseudocode in Fig. 2 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 [5]. Then we say that a given trend is, e.g., “decreasing to a degree 0.8”, if $\mu_{\text{decreasing}}(\alpha) = 0.8$, where $\mu_{\text{decreasing}}$ is the membership function of a fuzzy set representing “decreasing” that is a best match for angle $\alpha$. 
3.2 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 may be as in Fig. 4 where the time axis is divided into appropriate units.

![Figure 4](image)

*Figure 4.* An example of a 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.

3.3 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 \( \frac{\sum (x_i - \bar{x})^2}{n} \), 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 \( \frac{\sum |x_i - \bar{x}|}{n} \). 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 $\varepsilon$-approximation (cf. Section 2) 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(a). Equation (1) assures that the distance is definitely smaller than $\varepsilon$. 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 $\varepsilon$. 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 [5]. 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.

4 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 [16], [17]). This data set is numeric and usually large, not comprehensible in its original form by the human being. In Yager’s approach (cf. Yager [22], Kacprzyk and Yager [14], and Kacprzyk, Yager and Zadrożny [15]) 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) $T$ 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 $T$ 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
\[
T(\text{most of employees earn low salary}) = 0.7 
\]
(2)
or, in a richer (extended) form, including a qualifier (e.g. young), by
\[
T(\text{most of young employees earn low salary}) = 0.9 
\]
(3)

Thus, basically, the core of a linguistic summary is a *linguistically quantified proposition* in the sense of Zadeh [26] which, for (2), may be written as
\[
Qy's \text{ are } P 
\]
and for (3), may be written as
\[
QRy's \text{ are } P 
\]

Then, $T$, 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. [26]), or other interpretations of linguistic quantifiers. In the former case, the truth values (from $[0, 1]$) of (4) and (5) are calculated, respectively, as
\[
T(Qy's \text{ are } P) = \mu_Q \left( \frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right) 
\]
(6)
\[
T(QRy's \text{ are } P) = \mu_Q \left( \frac{\sum_{i=1}^{n} (\mu_R(y_i) \land \mu_P(y_i))}{\sum_{i=1}^{n} \mu_R(y_i)} \right) 
\]
(7)
where $\wedge$ is the minimum operation (more generally it can be another appropriate operation, notably a $t$-norm), and $Q$ is a fuzzy set representing the linguistic quantifier in the sense of Zadeh [26], i.e. $\mu_Q : [0, 1] \rightarrow [0, 1]$, $\mu_Q(x) \in [0, 1]$. We consider regular non-decreasing monotone quantifiers such that:

$$\mu(0) = 0, \quad \mu(1) = 1$$  \hspace{1cm} (8)

$$x_1 \leq x_2 \Rightarrow \mu_Q(x_1) \leq \mu_Q(x_2)$$  \hspace{1cm} (9)

They can be exemplified by “most” given as in (10):

$$\mu_Q(x) = \begin{cases} 
1 & \text{for } x \geq 0.8 \\
2x - 0.6 & \text{for } 0.3 < x < 0.8 \\
0 & \text{for } x \leq 0.3
\end{cases}$$  \hspace{1cm} (10)

5 Protoforms of linguistic trend summaries

It was shown by Kacprzyk and Zadrozny [16] that Zadeh’s [27] concept of a 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 the following protoforms of frequency based summaries:

- a protoform of a short form of linguistic summaries:

$$Q \text{ trends are } P$$  \hspace{1cm} (11)

and exemplified by:

Most of trends are of a large variability

- a protoform of an extended form of linguistic summaries:

$$QR \text{ trends are } P$$  \hspace{1cm} (12)

and exemplified by:

Most of slowly decreasing trends are of a large variability

These basic and intuitively appealing types of summaries can easily be extended to duration based summaries:
• a protoform of a short form of linguistic summaries:

\[ \text{Trends that took } Q \text{ time are } P \quad (13) \]

and exemplified by:

\[ \text{Trends that took most of the time are of a large variability} \]

• a protoform of an extended form of linguistic summaries:

\[ \text{R trends that took } Q \text{ time are } P \quad (14) \]

and exemplified by:

\[ \text{Slowly decreasing trends that took most of the time are of a large variability} \]

Since the interpretation of these extended forms may sometimes be difficult in many areas, for instance, in case of business data, we will not use them in this paper.

The truth values of the above types and forms of linguistic summaries will be found using the classic Zadeh’s calculus of linguistically quantified propositions as it is effective and efficient, and provides the best conceptual framework within which to consider a linguistic quantifier driven aggregation of partial trends that is the crucial element of our approach.

Using Zadeh’s [26] fuzzy logic based calculus of linguistically quantified propositions, a (proportional, nondecreasing) linguistic quantifier \( Q \) is assumed to be a fuzzy set defined in the unit interval \([0,1]\) as, e.g (10).

The truth values (from [0,1]) of (11) and (12) are calculated, respectively, as

\[ T(Qy’\text{s are } P) = \mu_Q \left( \frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right) \quad (15) \]

\[ T(QRy’\text{s are } P) = \mu_Q \left( \frac{\sum_{i=1}^{n} (\mu_R(y_i) \land \mu_P(y_i))}{\sum_{i=1}^{n} \mu_R(y_i)} \right) \quad (16) \]

where \( \land \) is the minimum operation. Both the fuzzy predicates \( P \) and \( R \) are assumed above to be of a rather simplified, atomic form referring to just one attribute. They can be extended to cover more sophisticated summaries involving some confluence of various, multiple attribute values as, e.g, “slowly decreasing and short”.

\[ (10) \]
6 Numerical experiments

The method proposed in this paper was tested on data coming from quotations of an investment fund that invests at most 50\% of assets in shares. Data shown in Figure 5 were collected from April 1998 until December 2006 with the value of one share equal to PLN 10.00 in the beginning of the period to PLN 45.10 at the end (PLN stands for the Polish Zloty). The minimal value recorded was PLN 6.88 while the maximal one during this period was PLN 45.15. The biggest daily increase was equal to PLN 0.91, while the biggest daily decrease was equal to PLN 2.41.

Using the Sklansky and Gonzalez algorithm and $\varepsilon = 0.25$ (PLN 0.25) we obtained 255 extracted trends. The shortest trend took 2 time units only, while the longest 71. The histogram of duration of trends is presented in Figure 6.

The histogram of variability of trends (in percents) is presented in Figure 8. Figure 7 shows the histogram of angles, which characterize dynamics of change.

Some interesting short form summaries obtained by using the method proposed, employing the classic Zadeh’s calculus of linguistically quantified propositions, and for different granulations of the dynamics of change, duration and variability, are:

- for 7 labels for the dynamics of change (quickly increasing, increasing, slowly increasing, constant, slowly decreasing, decreasing and quickly
decreasing), 5 labels for the duration (very long, long, medium, short, very short) and the variability (very high, high, medium, low, very low):

- Most trends are very short, $T = 0.78$
- Trends that took almost all of the time are constant, $T = 0.639$
- Most trends with a low variability are constant, $T = 0.974$
- Most slowly decreasing trends are of a very low variability, $T =$
Figure 8. Histogram of variability (IQR) of trends in %

0.636

- 5 labels for the dynamics of change (increasing, slowly increasing, constant, slowly decreasing, decreasing), 3 labels for the duration (short, medium, long) and 5 labels for the variability (very high, high, medium, low, very low):

- Trends that took most of the time are constant, $T = 0.692$
- Trends that took most of the time are of a medium length, $T = 0.506$
- Most of slowly increasing trends are of a medium length, $T = 0.798$
- Most of trends with a low variability are constant, $T = 0.567$
- Most of trends with a very low variability are short, $T = 0.909$
- Most trends with a high variability are of a medium length, $T = 0.801$
- More or less a half of medium length trends are constant, $T = 0.891$
- Almost none of trends with a very high variability are long, $T = 1$
- Almost none of decreasing trends are long, $T = 1$
- Almost none of increasing trends are long, $T = 1$
As it can be seen, the results obtained, that is the particular linguistic summaries and their associated truth values, are intuitively appealing while looking at the time series under consideration. In addition, these summaries have been found interesting by domain experts though a detailed analysis from the point of view of financial analyses is beyond the scope of this paper.

7 Concluding remarks

We proposed new types of linguistic summaries of time series. The derivation of a linguistic summary of a time series was related to a linguistic quantifier driven aggregation of trends, and we employed the classic Zadeh’s calculus of linguistically quantified propositions only with the classic minimum. We showed an application to the analysis of time series data on daily quotations of an investment fund over an eight year period, present some interesting linguistic summaries obtained, and showed results. The results are very promising.

References


