Engineering Applications of Artificial Intelligence 56 (2016) 273–280

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Using linguistic summaries and concepts for understanding large data

Ronald R. Yager[a,*], Rachel L. Yager[b]

[a] Machine Intelligence Institute, Iona College, New Rochelle, NY 10801, United States
[b] Department of Management, Metropolitan College of New York, New York, NY 10013, United States

ARTICLE INFO

Keywords:
Summarization
Linguistic Summaries
Concept Formation
Aggregation
Data Mining

ABSTRACT

We introduce the basic ideas of linguistic summaries. We describe the validation of the summary based upon the compatibility of the data and linguistic terms used in the summary. We next discuss the idea of concepts and describe how to build these concepts from an aggregation of constituent concepts. We discuss the role that hierarchies play in formulating rich concepts. We show how to include these concepts in linguistic summaries and in particular how to validate a summary that contains concepts. We look at the role of aggregation operators in the construction of the rich concepts. We explain how to build concepts using the OWA aggregation operator. Finally we discuss the use of the Choquet integral as means of building rich concepts.

1. Introduction

An important goal in mining large data is to provide the information contained in the data in the most understandable manner. Here we discuss some technologies that can enable us to attain this goal. Summarization provides one means of getting a global picture of a collection of data. The idea of linguistic summaries (Yager, 1989, 1991; Kacprzyk et al., 2000; Kacprzyk and Yager, 2001; Kacprzyk et al., 2001; Boran et al., 2014) is to enable the summarization of a collection of data using linguistic terms and thereby making the information provided more comprehensible. Furthermore, the focus of these summaries should involve terms and ideas that are relevant to the goals and objectives of the recipient of the summary. Our goal in this work is to provide for an extension of the original work on linguistic summaries by providing the ability to express the summaries in terms of complex concepts, which we refer to these as rich concepts. By using these concepts we are able expand the space of ideas about which we can make summaries from a collection of data. Here we have contributed to this goal by showing how we can build these rich concepts by an aggregation of simpler constituent concepts in a kind of hierarchical fashion. We also show how to validate these concepts based on the truth of the constituent components of the concept.

2. Linguistic summaries

In (Yager, 1989, 1991) Yager introduced the idea of linguistic summaries as a user-friendly method of summarizing information in a database. Kacprzyk and other researchers (Kacprzyk et al., 2000; Kacprzyk and Yager, 2001; Kacprzyk et al., 2001; Boran et al., 2014; Kacprzyk and Strykowski, 1999; Yager and Kacprzyk, 1999; Kacprzyk and Zadrozny, 2010; Boran et al., 2013; Wilbik and Keller, 2013) have made considerable use of the idea of linguistic summary. Here we briefly review some tools associated with linguistic summaries.

Assume we have a database Y ={y_1, ..., y_n} where the y_i are the objects in with database. Assume V is some attribute associated with the elements in the database having as its domain X. For example, if each y_i is a person then V could be their age. Here then for each y_i we have a value $V(y_i) = a_i$ where $a_i \in X$. Associated with the attribute V is a data set D =[a_1, ..., a_n], bag (Yager, 1986), containing the values of V assumed by the objects is the database Y. We emphasize that a bag or multi-set allows multiple elements with the same value.

A linguistic summary associated with V is a global statement based on the values in D. If V is the attribute age some examples of simple linguistic summaries are.

Most people in the database are about 30 year old
Few people in the database are old
Nearly a quarter of the people in the database are middle aged
Formally a simple linguistic summary is a statement of the form

Q objects in the database have V is **S**.

In the above ***S*** is called the summarizer and ***Q*** is called the quantity is agreement. Also associated with a linguistic summary is a measure of validity of the summary, τ. The value τ is used to indicate the truth of statement that **Q** objects have the property that V is ***S*** in the light of the data set D.

A fundamental characteristic of this formulation is that the summarizer and quantity in agreement are expressed in linguistic

[*] Corresponding author.
E-mail addresses: yager@panix.com (R.R. Yager), ryager@mcny.edu (R.L. Yager).

http://dx.doi.org/10.1016/j.engappai.2016.09.009
Received 5 April 2016; Received in revised form 1 September 2016; Accepted 21 September 2016
Available online 17 October 2016
0952-1976/

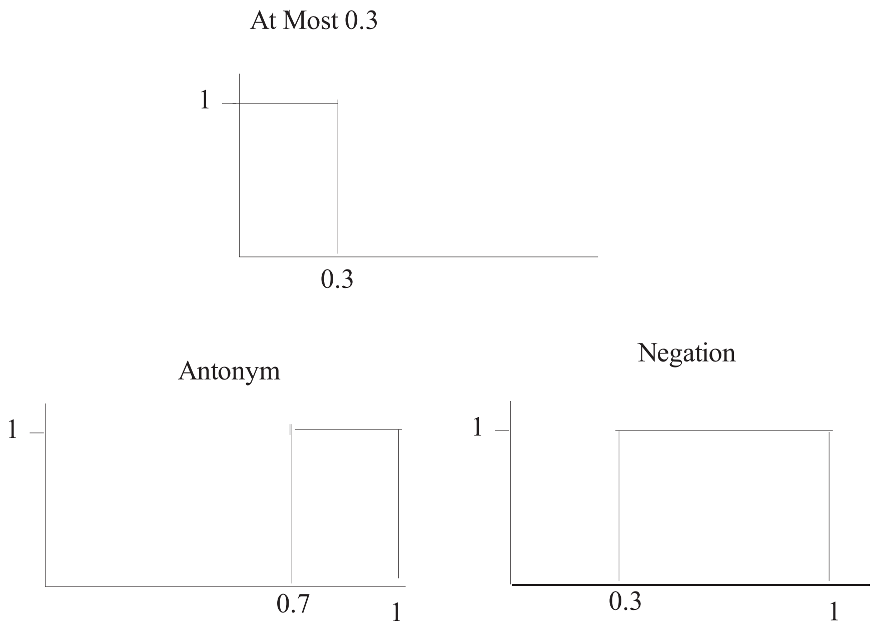

Fig. 1. Distinction between antonym and negation..

terms. One advantage of the use of linguistic summaries is that they provide statements about the dataset in terms that a very easy for people to comprehend. In (Yager, 2012) Yager showed that linguistic summarizes are closely related to what Zadeh has called Z-numbers (Zadeh, 2011).

Using fuzzy subsets we are able to provide a formal semantics for the terms used in a linguistic summary. In a procedure to be subsequently described, we shall use this ability to formalize the summarizers and quantity in agreement as fuzzy sets to enable us to evaluate the validity of a linguistic summary. This validation process will be based upon a determination of the compatibility of the linguistic summary with the data set D. It should be pointed out that for a given attribute we can conjecture numerous different summaries, then with the aid of the data set D we can obtain the validity, τ, of a proposed linguistic summary.

In developing our approach to validating a linguistic summary considerable use will be made of the ability to represent a linguistic summarizer by a fuzzy subset over the domain of the attribute. If V is some attribute taking its value from the domain X and if **S** is some concept associated with this attribute we can represent **S** by a fuzzy subset S on X such that for each $x \in X$, $S(x) \in [0,1]$ is the degree of compatibility of the value x with the concept **S**. If V is age and **S** is the concept middle age then S(40) indicates the degree to which 40 years old is compatible with the idea of middle age. Even in environments in which the underlying domain is non-numeric using this approach allows us to obtain numeric values for the membership grade in the fuzzy subset S corresponding to the concept **S**. For example if V is the attribute city of residence that takes as its domain the cities in the U.S. we can express the concept **S**, “lives near New York”, as a fuzzy subset. The second component in our linguistic summary is the quantity in agreement **Q**. These objects belong to a class of concepts called linguistic quantifiers (Zadeh, 1983). Examples of linguistic quantifiers are terms such as most, few, about half, all. Essentially linguistic quantifiers are fuzzy proportions, an alternative view of these subjects are generalized logical quantifiers. In (Zadeh, 1983) Zadeh suggested we could represent these linguistic quantifiers as fuzzy subsets of the unit interval. Using this representation the membership grade of any proportion $r \in [0,1]$ in the fuzzy set Q corresponding to the linguistic quantifier **Q**, Q(r), is a measure of the compatibility of the proportion r with the linguistic quantifier we are representing by the fuzzy subset Q. For example if Q is the fuzzy set corresponding to the quantifier *Most* then Q(0.9) represents the degree to which the proportion 0.9 satisfies the concept *Most*.

In (Yager, 1985) Yager identified three classes of linguistic quantifiers that cover most of those used in natural language. (1) Q is said to be monotonically non-decreasing if $r_1 >\ _2 \Rightarrow Q(r_1) \geq Q(r_2)$, examples of this type of quantifier are *at lest 30%*, *most*, *all*. (2) A quantifier Q is said to monotonically non-increasing if $r_1 > r_2 \Rightarrow Q(r_1) \leq Q(r_2)$, examples of this type of quantifiers are *at most 30%, few*, *none*. (3) A quantifier Q is said to be unimodal if there exists two values $a \leq b$ both contained in the unit interval such that for $r < a$, Q is monotonically non-decreasing, for $r > b$, Q is monotonically non-increasing and for $r \in [a, b]$, $Q(r) = 1$, an example of this type of quantifier is *about 0.3*.

An important idea that can be associated with a linguistic quantifier is the concept of an antonym. If Q is a linguistic quantifier its antonym is also a linguistic quantifier, denoted $\hat{Q}$, such that $\hat{Q}(r) = Q(1 - r)$. The operation of taking an antonym is involutionary, that is $\hat{\hat{Q}} = Q$. From this we see that antonyms come in pairs. Prototypical examples of antonym pairs are **all-none** and **few-many**. Consider the quantifier *at most 0.3* defined as $Q(r) = 1$ if $r \leq 0.3$ and $Q(r) = 0$ if $r > 0.3$. Its antonym has $\hat{Q}(r) = 1$ if $r \geq 0.3$ and $\hat{Q}(1 - r) = 0$ if $r \geq 0.3$. This can be seen to be equivalent to $\hat{Q}(r) = 1$ if $r \geq 0.7$ and $\hat{Q}(r) = 0$ if $r < 0.7$. Thus the antonym of *at most 0.3* is *at least 0.7*.

Care must be taken to distinguish between the antonym of a quantifier and its negation. We recall the negation of Q denoted $\overline{Q}$ is defined such that $\overline{Q}(r) = 1 - Q(r)$. We see that the negation of ***at most 0.3*** is $\overline{Q}(r) = 0$ if $r \leq 0.3$ and $\overline{Q}(r) = 1$ if $r \geq 0.3$, this corresponds ***to at least 0.3"***. In Fig. 1 we plot these different quantifiers related to *at most 0.3*.

Having discussed the concepts of summarizer and quantity in agreement we are now in a position to describe the methodology used to calculate the validity **τ** of a linguistic summary. Assume $D = [a_1, a_2, ..., a_n]$ is the collection of values that appear in the database for the attribute V. Consider the linguistic summary:

Q items in the database have values for V that are S.

The basic procedure to obtain the validity **τ** of this summary in the face of the data D is:

(1) For each a_i in D, calculate $S(a_i)$, the degree to which a_i satisfies the summarizer S.
(2) Let r $= \frac{1}{n}\sum_{i=1}^{n} S(a_i)$, the proportion of D that satisfy S.