Optimal Selection of Proportional Bounding Quantifiers in Linguistic Data Summarization

Ingo Glöckner

FernUniversität in Hagen, 58084 Hagen, Germany, iglockner@web.de

Summary. Proportional bounding quantifiers like "Between p_1 and p_2 percent" are potentially useful for expressing linguistic summaries of data. Given p_1, p_2, existing methods for data summarization based on fuzzy quantifiers can be used to assign a quality score to the summary. However, the problem remains how the optimal choice of p_1, p_2 in the range $0 \leq p_1 \leq p_2 \leq 100\%$ can be established. Moreover, the proposed quality indicators are rather heuristic in nature. The paper presents a method for computing the optimal bounding quantifier which best summarizes the given data. Specifically, the most specific quantifier will be chosen which results in the highest validity score of the summary given a constraint on the the percentage range $p_2 - p_1$. The method not only assigns validity scores to the quantifiers of interest but also determines the best choice of quantifier in $\mathcal{O}(N \log m)$ time, where N is the size of the base set and m the number of different membership grades in the fuzzy arguments.

1 Introduction

A framework for generating linguistic summaries from imprecise data and for evaluating their usefulness has been developed by Yager [8, 10, 7] and refined by Kacprzyk and Strykowski [6]. We assume a finite base set $E \neq \varnothing$ of individuals of interest, and a description of these individuals by fuzzy sets $X \in \widetilde{\mathscr{P}}(E)$. In practice, the data will likely be described by a set A of attributes $a : E \longrightarrow V_a$ which assign an attribute value $a(e)$ to each individual. The fuzzy sets on E, in turn, will only indirectly be given by fuzzy sets declared on the attribute values. Thus, a fuzzy set $X' \in \widetilde{\mathscr{P}}(V_a)$ declared on the attribute range of $a \in A$ gives rise to the associated fuzzy set $X \in \widetilde{\mathscr{P}}(E)$ defined by $\mu_X(e) = \mu_{X'}(a(e))$. The attribute-based description of the data in a database will not be of relevance in this paper, however, so that we drop it for simplicity. From this simplified viewpoint, then, a linguistic summary has the form "Q objects are X's" or "Q X_1's are X_2's" (with an associated 'validity' or 'truthfulness' score $\tau \in [0,1]$). $Q \in \{\textit{almost all}, \textit{many}, \ldots\}$ is called the 'quantity in agreement' of the summary.

While existing data summarization systems are mostly based on Zadeh's Σ-count or FG-count approach [11] or on Yager's OWA operators [9], we will assume a broader framework which avails us with a uniform analysis of all kinds of linguistic

Table 1. Main types of linguistic quantifiers

Type	example	definition
absolute unrestrictive	There are more than 3 Y's	$Q(Y) = q(\|Y\|)$
absolute	More than 3 Y_1's are Y_2's	$Q(Y_1,Y_2) = q(\|Y_1 \cap Y_2\|)$
exception	All except 3 Y_1's are Y_2's	$Q(Y_1,Y_2) = q(\|Y_1 \setminus Y_2\|)$
proportional	Two of three Y_1's are Y_2's	$Q(Y_1,Y_2) = \begin{cases} f(\frac{\|Y_1 \cap Y_2\|}{\|Y_1\|}) & \|Y_1\| > 0 \\ v_0 & \text{else} \end{cases}$
cardinal comparative	More Y_1's than Y_2 are Y_3	$Q(Y_1,Y_2,Y_3) = q(\|Y_1 \cap Y_3\|, \|Y_2 \cap Y_3\|)$

quantifiers [3, 5, 4]. This framework is inspired by the linguistic Theory of Generalized Quantifiers (TGQ) [1]. It extends the notion of a (two-valued) generalized quantifier to fuzzy arguments and gradual quantification results in the obvious way:

Definition 1. *An nary* fuzzy quantifier *on a base set $E \neq \varnothing$ is a mapping $\widetilde{Q} : \widetilde{\mathscr{P}}(E)^n \longrightarrow [0,1]$ (E needs not be finite).*

Expressing an NL quantifier of interest in terms of a fuzzy quantifier is not an easy task, though – mainly because a simple cardinality-based definition is no longer possible when the arguments of the quantifier are fuzzy. We therefore introduce so-called semi-fuzzy quantifiers which serve as a simplified description of the target quantifier.

Definition 2. *An nary* semi-fuzzy quantifier *on a base set $E \neq \varnothing$ is a mapping $Q : \mathscr{P}(E)^n \longrightarrow [0,1]$.*

Semi-fuzzy quantifiers are easier to define than fuzzy quantifiers because one needs not describe their interpretation for fuzzy arguments. The specification of an NL quantifier in terms of a semi-fuzzy quantifier will be linked to the target fuzzy quantifier (which also accepts fuzzy arguments) by a quantifier fuzzification mechanism.

Definition 3. *A* quantifier fuzzification mechanism *(QFM) $\mathscr{F}$ assigns a fuzzy quantifier $\mathscr{F}(Q) : \widetilde{\mathscr{P}}(E)^n \longrightarrow [0,1]$ to each semi-fuzzy quantifier $Q : \mathscr{P}(E)^n \longrightarrow [0,1]$.*

Table 1 lists the main types of semi-fuzzy quantifiers which are also of interest for data summarization. In this paper, we will restrict attention to *proportional bounded quantifiers*, i.e. quantifiers defined by

$$rate_{[r_1,r_2]}(Y_1,Y_2) = \begin{cases} 1 & : \quad Y_1 \neq \varnothing \wedge \frac{|Y_1 \cap Y_2|}{|Y_1|} \in [r_1,r_2] \\ 0 & : \quad \text{else} \end{cases}$$

These quantifiers are of apparent utility for expressing summaries like "between p_1 and p_2 percent of the X_1's are X_2's" ($r_1 = p_1/100$, $r_2 = p_2/100$) or "at least p percent of the X_1's are X_2's" ($r_1 = p/100$, $r_2 = 1$).

From the perspective of natural language use (i.e. pragmatics), truthfulness of $Q(Y_1,Y_2)$ only indicates a possible use of Q – which might represent a very unusual case of applying Q, however. In other words, the truth score $\tau = \mathscr{F}(Q)(X_1,X_2)$

which judges the validity of the summary "$Q\ X_1$'s are X_2's", is not restrictive enough to guide quantifier selection to the most appropriate choice of Q. Consider *at least eighty percent*, for example. Clearly the corresponding quantifier should also be true if *all* X_1's are X_2's. However, only the quantifier *all* is appropriate for describing this situation, while *at least eighty percent* has a very low appropriateness grade in this case. Existing approaches to linguistic data summarization have introduced various quality indicators for quantifier selection to solve this problem. While Yager [10] uses only the validity score and a metric for informativeness, Kacprzyk and Strykowski [6] use a multi-dimensional measure based on the degree of truth, the degree of imprecision, the degree of covering, the degree of 'appropriateness' and the length of the summary. The proposed quality indicators are rather heuristic in nature, though, while in this paper, we target at a more principled solution.

Hence let $X_1, X_2 \in \widetilde{\mathscr{P}}(E)$ be given. We further assume a fixed model of fuzzy quantification, i.e. a QFM $\mathscr{F}$ (see below). The summary generation is controlled by a constraint $\delta_{\max} \in [0,1)$ which specifies the maximal admissible percentage range $r_2 - r_1$ (a summary "between 20% and 90% of the X_1's are X_2's" might be pretty odd, for example). We search for an optimal choice of $Q^* = \mathit{rate}_{[r_1^*, r_2^*]}$, $r_1^* \leq r_2^*$, with

1. $r_2^* - r_1^* \leq \delta_{\max}$.
2. For all $r_1' \leq r_2'$ with $r_2' - r_1' \leq \delta_{\max}$, $\mathscr{F}(\mathit{rate}_{[r_1', r_2']})(X_1, X_2) \leq \mathscr{F}(Q^*)(X_1, X_2)$.
3. If $\mathscr{F}(\mathit{rate}_{[r_1', r_2']})(X_1, X_2) = \mathscr{F}(Q^*)(X_1, X_2)$, then $r_1' \leq r_1^*$ and $r_2^* \leq r_2'$.

The first condition asserts that the percentage span $r_2^* - r_1^*$ of the optimal quantifier is within the limits given by $\delta_{\max}$, i.e. the summary is sufficiently specific. The second condition asserts that Q^* is an optimal choice of such a quantifier, which results in the highest validity score. The third condition ensures that Q^* is the narrowest choice of all quantifiers optimal with respect to the validity score. Let me now explain how Q^* can actually be computed. In order to simplify presentation, we will further assume that $\mathscr{F}(Q^*)(X_1, X_2) > \frac{1}{2}$, i.e. summaries with low truth scores will be ignored.

2 Choosing the model $\mathscr{F}$

We need a concrete, well-motivated choice of $\mathscr{F}$ to determine the validity score $\tau = \mathscr{F}(Q)(X_1, X_2)$ of the summary. The following construction will assign a suitable choice of fuzzy connectives to the given QFM.

Definition 4. *Let $\mathscr{F}$ be a QFM and $f : \{0,1\}^n \longrightarrow [0,1]$ a (semi-fuzzy) truth function. The* induced fuzzy truth function $\widetilde{\mathscr{F}}(f) : [0,1]^n \longrightarrow [0,1]$ *is defined by $\widetilde{\mathscr{F}}(f) = \mathscr{F}(f \circ \eta^{-1}) \circ \widetilde{\eta}$, where $\eta : \{0,1\}^n \longrightarrow \mathscr{P}(\{1,\ldots,n\})$ and $\widetilde{\eta} : [0,1]^n \longrightarrow \widetilde{\mathscr{P}}(\{1,\ldots,n\})$ are defined by $\eta(y_1,\ldots,y_n) = \{i : y_i = 1\}$ and $\mu_{\widetilde{\eta}(x_1,\ldots,x_n)}(i) = x_i$, respectively.*

The fuzzy set operations $\widetilde{\cup} : \widetilde{\mathscr{P}}(E)^2 \longrightarrow \widetilde{\mathscr{P}}(E)$ (fuzzy union) and $\widetilde{\neg} : \widetilde{\mathscr{P}}(E) \longrightarrow \widetilde{\mathscr{P}}(E)$ (fuzzy complement) will be defined element-wise in terms of fuzzy disjunction $\widetilde{\vee} = \widetilde{\mathscr{F}}(\vee)$ and fuzzy negation $\widetilde{\neg} = \widetilde{\mathscr{F}}(\neg)$. Based on these induced fuzzy connectives and fuzzy set operations, we can define a class of well-behaved models.

Table 2. Known classes of standard models: an overview

Type	Construction
$\mathscr{F}_\Omega$-DFS	From supervaluation results of three-valued cuts:
	$X_\gamma^{\min} = \begin{cases} X_{\geq \frac{1}{2}+\frac{1}{2}\gamma} & \gamma \in (0,1] \\ X_{>\frac{1}{2}} & \gamma = 0 \end{cases}$ $X_\gamma^{\max} = \begin{cases} X_{>\frac{1}{2}-\frac{1}{2}\gamma} & \gamma \in (0,1] \\ X_{\geq\frac{1}{2}} & \gamma = 0 \end{cases}$
	$\mathscr{T}_\gamma(X_i) = \{Y : X_\gamma^{\min} \subseteq Y \subseteq X_\gamma^{\max}\}$
	$S_{Q,X_1,\ldots,X_n}(\gamma) = \{Q(Y_1,\ldots,Y_n) : Y_i \in \mathscr{T}_\gamma(X_i), i = 1,\ldots,n\}$
	$\mathscr{F}_\Omega(Q)(X_1,\ldots,X_n) = \Omega(S_{Q,X_1,\ldots,X_n})$
$\mathscr{F}_\xi$-DFS	From suprema and infima of supervaluations:
	$\top_{Q,X_1,\ldots,X_n}(\gamma) = \sup S_{Q,X_1,\ldots,X_n}(\gamma)$ $\bot_{Q,X_1,\ldots,X_n}(\gamma) = \inf S_{Q,X_1,\ldots,X_n}(\gamma)$
	$\mathscr{F}_\xi(Q)(X_1,\ldots,X_n) = \xi(\top_{Q,X_1,\ldots,X_n}, \bot_{Q,X_1,\ldots,X_n})$
$\mathscr{M}_\mathscr{B}$-DFS	From fuzzy median of supervaluation results:
	$Q_\gamma(X_1,\ldots,X_n) = \mathrm{med}_{1/2}(\top_{Q,X_1,\ldots,X_n}(\gamma), \bot_{Q,X_1,\ldots,X_n}(\gamma))$
	$\mathscr{M}_\mathscr{B}(Q)(X_1,\ldots,X_n) = \mathscr{B}(Q_\gamma(X_1,\ldots,X_n)_{\gamma\in[0,1]})$

Definition 5. *A QFM $\mathscr{F}$ is called a* determiner fuzzification scheme *(DFS) if it satisfies the following conditions for all semi-fuzzy quantifiers $Q : \mathscr{P}(E)^n \longrightarrow [0,1]$ and fuzzy arguments $X_1,\ldots,X_n \in \widetilde{\mathscr{P}}(E)$:*

(a) $\mathscr{F}(Q) = Q$ *if* $n = 0$*;*
(b) $\mathscr{F}(Q)(Y) = Q(Y)$ *for crisp* $Y \in \mathscr{P}(E)$*,* $n = 1$*;*
(c) $\mathscr{F}(\pi_e) = \widetilde{\pi}_e$ *for all* $E \neq \varnothing$*,* $e \in E$*, where* $\pi_e(Y) = 1$ *iff* $e \in Y$ *and* $\widetilde{\pi}_e(X) = \mu_X(e)$*;*
(d) $\mathscr{F}(Q')(X_1,\ldots,X_n) = \widetilde{\neg}\mathscr{F}(Q)(X_1,\ldots,X_{n-1},\widetilde{\neg}X_n)$ *whenever the semi-fuzzy quantifier Q' is defined by $Q'(Y_1,\ldots,Y_n) = \widetilde{\neg}Q(Y_1,\ldots,Y_{n-1},\neg Y_n)$ for all crisp Y_i;*
(e) $\mathscr{F}(Q')(X_1,\ldots,X_{n+1}) = \mathscr{F}(Q)(X_1,\ldots,X_{n-1},X_n \,\widetilde{\cup}\, X_{n+1})$ *whenever Q' is defined by $Q'(Y_1,\ldots,Y_{n+1}) = Q(Y_1,\ldots,Y_{n-1},Y_n \cup Y_{n+1})$ for all crisp Y_i;*
(f) $\mathscr{F}(Q)(X_1,\ldots,X_n) \geq \mathscr{F}(Q)(X_1,\ldots,X_{n-1},X_n')$ *if* $X_n \subseteq X_n'$ *given that* $Q(Y_1,\ldots,Y_n) \geq Q(Y_1,\ldots,Y_{n-1},Y_n')$ *for all crisp* Y_i*,* $Y_n \subseteq Y_n'$*;*
(g) $\mathscr{F}(Q \circ \times_{i=1}^n \widehat{\mathscr{F}}(f_i)) = \mathscr{F}(Q) \circ \times_{i=1}^n \widehat{f_i}$ *for all* $f_i : E' \longrightarrow E$*, where* $\widehat{f}(Y) = \{f(e) : e \in Y\}$ *for all crisp Y and* $\mu_{\widehat{\mathscr{F}}(f)(X)}(e) = \mathscr{F}(\pi_e \circ \widehat{f})(X)$*.*

The choice of postulates (a) through (g) was based on a large catalogue of semantic desiderata from which a minimal (independent) system of core requirements was then distilled. The total list of desiderata validated by these models is discussed in [4]. A DFS will be called a *standard DFS* if it induces the standard set of fuzzy connectives min, max and $1-x$.

Table 2 lists three general constructions of models which result in the classes of $\mathscr{F}_\Omega$, $\mathscr{F}_\xi$ and $\mathscr{M}_\mathscr{B}$ DFSes.[1] The $\mathscr{F}_\Omega$-DFSes form the broadest class of standard DFSes currently known. All practical $\mathscr{F}_\Omega$-DFSes belong to the more regular $\mathscr{F}_\xi$ class, though. The $\mathscr{M}_\mathscr{B}$-DFSes can be characterized as the subclass of those $\mathscr{F}_\xi$ models which propagate fuzziness (in the sense of being compatible with a natural

[1] Here, $X_{\geq\alpha}$ denotes the α-cut and $X_{>\alpha}$ the strict α-cut, respectively. Moreover, $\mathrm{med}_{1/2}(x,y)$ is the fuzzy median, i.e. the second-largest of the three values x, y, $\frac{1}{2}$.

fuzziness order). The most prominent example is the following standard DFS $\mathscr{M}_{\mathrm{CX}}$, which generalizes the Zadeh's FG-count approach:

$$\mathscr{M}_{\mathrm{CX}}(Q)(X_1,\ldots,X_n) = \begin{cases} \frac{1}{2}+\frac{1}{2}\sup\{\gamma : \bot(\gamma) \geq \frac{1}{2}+\frac{1}{2}\gamma\} & : \quad \bot(0) > \frac{1}{2} \\ \frac{1}{2}-\frac{1}{2}\sup\{\gamma : \top(\gamma) \leq \frac{1}{2}-\frac{1}{2}\gamma\} & : \quad \top(0) < \frac{1}{2} \\ \frac{1}{2} & : \quad \text{else} \end{cases}$$

abbreviating $\top(\gamma) = \top_{Q,X_1,\ldots,X_n}(\gamma)$ and $\bot(\gamma) = \bot_{Q,X_1,\ldots,X_n}(\gamma)$. In the following, it is sufficient to consider the model $\mathscr{F} = \mathscr{M}_{\mathrm{CX}}$ only because all standard DFSes coincide with $\mathscr{M}_{\mathrm{CX}}$ for two-valued quantifiers, and because proportional bounding quantifiers are two-valued quantifiers.

3 Implementation of proportional bounding quantifiers

In order to describe the method for optimal quantifier selection, we must recall the computational analysis of quantitative quantifiers in $\mathscr{F}_\xi$ models developed in [5]. In the following, we assume a finite base set $E \neq \varnothing$ of cardinality $|E| = N$. For given fuzzy arguments $X_1,\ldots,X_n \in \widetilde{\mathscr{P}}(E)$, the set of relevant cutting levels is given by $\Gamma(X_1,\ldots,X_n) = \{2\mu_{X_i}(e)-1 : \mu_{X_i}(e) \geq \frac{1}{2}\} \cup \{1-2\mu_{X_i}(e) : \mu_{X_i}(e) < \frac{1}{2}\} \cup \{0,1\}$. The computation of quantifiers will be based on an ascending sequence of cutting levels $0 = \gamma_0 < \gamma_1 \cdots < \gamma_{m-1} < \gamma_m = 1$ with $\{\gamma_1,\ldots,\gamma_m\} \supseteq \Gamma(X_1,\ldots,X_n)$ (usually we will have an equality here). For $\gamma = 0,\ldots,m-1$, we abbreviate $\overline{\gamma}_j = \frac{\gamma_j+\gamma_{j+1}}{2}$. We further let $\top_j = \top_{Q,X_1,\ldots,X_n}(\overline{\gamma}_j)$ and $\bot_j = \bot_{Q,X_1,\ldots,X_n}(\overline{\gamma}_j)$. As a prerequisite for implementing the quantifier selection, let us rewrite $\mathscr{M}_{\mathrm{CX}}$ as a function of the finite sample $\Gamma(X_1,\ldots,X_n)$ of (three-valued) cut levels.

Proposition 1. *Let $Q : \mathscr{P}(E)^n \longrightarrow [0,1]$, $X_1,\ldots,X_n \in \widetilde{\mathscr{P}}(E)$ and $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_{m-1} < \gamma_m = 1$ be given, $\Gamma(X_1,\ldots,X_n) \subseteq \{\gamma_0,\ldots,\gamma_m\}$. For $j \in \{0,\ldots,m-1\}$ let $B_j = 2\bot_j - 1$ if $\bot_0 \geq \frac{1}{2}$ and $B_j = 1-2\top_j$ otherwise. Further let*

$$\widehat{J} = \{j \in \{0,\ldots,m-1\} : B_j \leq \gamma_{j+1}\}, \quad \widehat{j} = \min \widehat{J}.$$

Then

$$\mathscr{M}_{\mathrm{CX}}(Q)(X_1,\ldots,X_n) = \begin{cases} \frac{1}{2}+\frac{1}{2}\max(\gamma_{\widehat{j}}, B_{\widehat{j}}) & : \quad \bot_0 > \frac{1}{2} \\ \frac{1}{2}-\frac{1}{2}\max(\gamma_{\widehat{j}}, B_{\widehat{j}}) & : \quad \top_0 < \frac{1}{2} \\ \frac{1}{2} & : \quad \textit{else.} \end{cases}$$

These formulas enable us to evaluate fuzzy quantifications in $\mathscr{M}_{\mathrm{CX}}$ (and for two-valued quantifiers like $\mathit{rate}_{[r_1,r_2]}$, in arbitrary standard DFSes). The computation of $\top_j$ and $\bot_j$ must be optimized, though, because a naive implementation which considers each $Y_i \in \mathscr{T}_{\overline{\gamma}_j}(X_i)$ will not achieve sufficient performance. To this end, we observe that for proportional bounding quantifiers, $\top_j$ and $\bot_j$ can be rewritten as

$$\top_j = \max\{q(c_1,c_2) : (c_1,c_2) \in R_j\} \qquad \bot_j = \min\{q(c_1,c_2) : (c_1,c_2) \in R_j\}$$

$$R_j = \{(|Y_1|, |Y_1 \cap Y_2|) : Y_i \in \mathscr{T}_\gamma(Y_i)\} \quad q(c_1,c_2) = \begin{cases} 1 & : \quad c_1 > 0 \wedge \frac{c_2}{c_1} \in [r_1, r_2] \\ 0 & : \quad \text{else.} \end{cases}$$

In numeric terms, the relation R_j can be precisely described as follows:

$$R_j = \{(c_1,c_2) : \ell_1 \leq c_1 \leq u_1, \max(\ell_2, c_1 - u_3) \leq c_2 \leq \min(u_2, c_1 - \ell_3)\}, \quad (1)$$

where $\ell_s = |Z_s|_\gamma^{\min} = |(Z_s)_\gamma^{\min}|$ and $u_s = |Z_s|_\gamma^{\max} = |(Z_s)_\gamma^{\max}|$, $\gamma = \overline{\gamma}_j$, depend on $Z_1 = X_1$, $Z_2 = X_1 \cap X_2$ and $Z_3 = X_1 \cap \neg X_2$, assuming the standard fuzzy intersection and complement.[2] We conclude that

$$\begin{aligned} \top_j &= \max\{q(c_1,c_2) : \ell_1 \leq c_1 \leq u_1, \max(\ell_2, c_1 - u_3) \leq c_2 \leq \min(u_2, c_1 - \ell_3)\} \\ &= \max\{q'(c_1) : \ell_1 \leq c_1 \leq u_1\} \\ \bot_j &= \min\{q(c_1,c_2) : \ell_1 \leq c_1 \leq u_1, \max(\ell_2, c_1 - u_3) \leq c_2 \leq \min(u_2, c_1 - \ell_3)\} \\ &= \begin{cases} 1 & : \quad \ell_1 > 0 \wedge r_1 \leq \frac{\ell_2}{\ell_2 + u_3} \wedge \frac{u_2}{u_2 + \ell_3} \leq r_2 \\ 0 & : \quad \text{else} \end{cases} \end{aligned}$$

where

$$q'(c_1) = \begin{cases} q(c_1, \min(u_2, c_1 - \ell_3)) & : \quad \min(u_2, c_1 - \ell_3) < r_1 \\ q(c_1, \max(\ell_2, c_1 - u_3)) & : \quad \max(\ell_2, c_1 - u_3) > r_1 \\ q(c_1, r_1) & : \quad \text{else.} \end{cases}$$

In order to compute a quantification result based on this formula, one must consider every choice of j (i.e. m cutting levels) and (at worst) $N = |E|$ choices of c_1. Thus, the complexity of evaluating a proportional bounding quantifier is $\mathscr{O}(Nm)$.

4 Optimal quantifier selection

Given $X_1, X_2 \in \widetilde{\mathscr{P}}(E)$, we define rate bound mappings $r^{\min}, r^{\max} : [0,1] \longrightarrow [0,1]$:

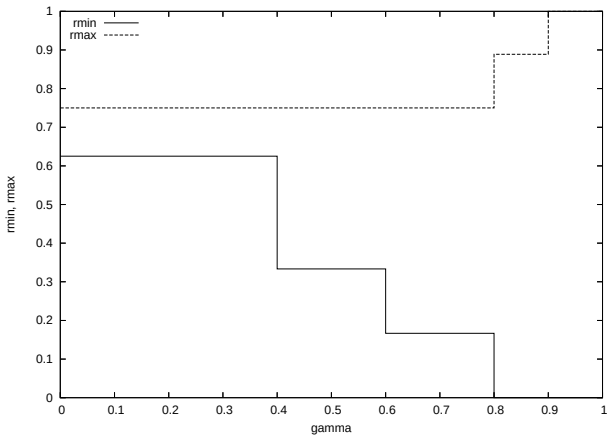
$$r^{\min}(\gamma) = \begin{cases} \frac{\ell_2}{\ell_2 + u_3} & : \quad \ell_1 > 0 \\ 0 & : \quad \ell_1 = 0 \end{cases} \qquad r^{\max}(\gamma) = \begin{cases} \frac{u_2}{u_2 + \ell_3} & : \quad \ell_1 > 0 \\ 1 & : \quad \ell_1 = 0 \end{cases}$$

where $\ell_s = \ell_s(j)$ and $u_s = u_s(j)$ are defined as (1), and $j = \max\{j : \gamma_j \leq \gamma\}$ for $\{\gamma_0, \ldots, \gamma_m\} = \Gamma(X_1, \ldots, X_n)$, $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_{m-1} < \gamma_m = 1$. Abbreviating $r_j^{\min} = r^{\min}(\overline{\gamma}_j)$, $r_j^{\max} = r^{\max}(\overline{\gamma}_j)$, it is easily shown from the above analysis of $\top_j$ and $\bot_j$ that for a proportional bounding quantifier $\mathit{rate}_{[r_1,r_2]}$ with $r_2 - r_1 \in [0,1)$,

$$\bot_j = \begin{cases} 1 & : \quad r_1 \leq r_j^{\min} \wedge r_j^{\max} \leq r_2 \\ 0 & : \quad \text{else} \end{cases} \quad (2)$$

In particular, choosing $r_1 = r_j^{\min}$ and $r_2 = r_j^{\max}$ will result in $\bot_j = 1 > \frac{1}{2} + \frac{1}{2}\overline{\gamma}_j$, i.e. $\mathscr{M}_{\mathrm{CX}}(\mathit{rate}_{[r_j^{\min}, r_j^{\max}]})(X_1, X_2) \geq \frac{1}{2} + \frac{1}{2}\gamma_{j+1} > \frac{1}{2}$. Hence let $j_* = \max\{j = 0, \ldots, m-1 :$

[2] The efficient computation of $\ell_s(j)$ and $u_s(j)$ from the histogram of Z_s is explained in [4, 5].



$\delta_{\max}$	$r_*^{\min}$	$r_*^{\max}$	τ
0.1	–	–	–
0.2	0.625	0.75	0.7
0.5	$0.\overline{3}$	0.75	0.8

Fig. 1. Example plot of $r^{\min}$ and $r^{\max}$ and results of quantifier selection

$r_j^{\max} - r_j^{\min} \leq \delta_{\max}\}$, $r_1^* = r_{j_*}^{\min}$ and $r_2^* = r_{j_*}^{\max}$. If $r_2^* - r_1^* > \delta_{\max}$, then no summarization based on bounding quantifiers is possible because X_1, X_2 are too fuzzy. In the normal case that $r_2^* - r_1^* \leq \delta_{\max}$, however, the choice of r_1^* and r_2^* will be optimal. To see this, suppose that $\mathscr{M}_{\mathrm{CX}}(\mathit{rate}_{[r_1,r_2]})(X_1,X_2) \geq \mathscr{M}_{\mathrm{CX}}(Q^*)(X_1,X_2)$, $Q^* = \mathit{rate}_{[r_1^*,r_2^*]}$. Because $\mathscr{M}_{\mathrm{CX}}(Q^*)(X_1,X_2) > \frac{1}{2}$, we also have $\mathscr{M}_{\mathrm{CX}}(Q')(X_1,X_2) > \frac{1}{2}$, where $Q' = \mathit{rate}_{[r_1,r_2]}$. Now consider $\perp^* = \perp_{Q^*,X_1,X_2}$ and $\perp' = \perp_{Q',X_1,X_2}$. Since Q^* and Q' are two-valued, $\perp^*$ and $\perp'$ are also two-valued. Moreover $\perp^*$ and $\perp'$ are monotonically non-increasing. Thus $\mathscr{M}_{\mathrm{CX}}(\mathit{rate}_{[r_1,r_2]})(X_1,X_2) \geq \mathscr{M}_{\mathrm{CX}}(Q^*)(X_1,X_2)$ is only possible if $\perp' \geq \perp^*$, since $\mathscr{M}_{\mathrm{CX}}$ preserves inequations of $\perp'$ and $\perp^*$. We see from $\perp' \geq \perp^*$ that $\perp'_{j_*} \geq \perp^*_{j_*} = 1$, i.e. $r_1 \leq r_1^*$ and $r_2^* \leq r_2$. Now suppose that $\mathscr{M}_{\mathrm{CX}}(Q')(X_1,X_2) > \mathscr{M}_{\mathrm{CX}}(Q^*)(X_1,X_2)$. Then $\mathscr{M}_{\mathrm{CX}}(Q') = \frac{1}{2} + \frac{1}{2}\gamma'$ for $\gamma' > \gamma_{j_*+1}$. In particular, $\perp'_{j_*+1} = 1$, i.e. $r_1 \leq r_{j_*+1}^{\min}$ and $r_{j_*+1}^{\max} \leq r_2$. By definition of j_*, $r_2 - r_1 \geq r_{j_*+1}^{\max} - r_{j_*+1}^{\min} > \delta_{\max}$, i.e. Q' exceeds the $\delta_{\max}$ limit. Thus in fact $\mathscr{M}_{\mathrm{CX}}(Q')(X_1,X_2) = \mathscr{M}_{\mathrm{CX}}(Q^*)(X_1,X_2)$ and $r_1 \leq r_1^*$, $r_2^* \leq r_2$, confirming the optimality of Q^*.

Apparently, the result can be computed in at most m steps ($j = 0,\ldots,m-1$). Computation of $\perp_j$ given j is possible in constant time, see (2). Therefore, the optimal choice of r_1 and r_2 can be computed in $\mathscr{O}(m)$ time. However, the computation of the cardinality coefficients must also be considered. The method for determining $\ell_s(j)$ and $u_s(j)$ described in [4, 5] requires a pre-computation of the histograms of X_1, X_2, which has complexity $\mathscr{O}(N \log m)$. In practice, this is sufficient to compute optimal bounding quantifiers even for large base sets.

Finally we consider an example of quantifier selection. Let us assume that $E = \{a,b,c,d,e,f,g,h,i\}$, $X_1 = 1/a + 0.9/b + 0.8/c + 0.8/d + 0.7/e + 0.7/f + 0.1/g + 1/h + 0.9/i$, $X_2 = 0.05/a + 0.9/b + 0.7/c + 1/d + 1/e + 0.8/f + 0.1/g + 0.5/h + 0.1/i$. The corresponding minimum and maximum rates and the results of the quantifier selection method for several choices of $\delta_{\max}$ are shown in Figure 1.

Here, $\tau = \mathscr{M}_{\mathrm{CX}}(Q^*)(X_1, X_2)$ is the resulting validity score of the summary. Notice that a summary like "Between 62.5% and 75% of the X_1's are X_2's" might be misleading because it gives an illusion of precision which is not justified by the data. We therefore suggest a subsequent *linguistic fitting* of the selected quantifier which should be adapted to a reasonable granularity scale. For example, assuming a 5% granularity scaling, the result for $\delta_{\max} = 0.2$ should better be expressed as "Between 60% and 75% of the X_1's are X_2's".

5 Conclusions

This paper was concerned with the problem of quantifier selection in fuzzy data summarization. For the important class of proportional bounding quantifiers, we have presented an efficient algorithm which computes the optimal quantifier in $\mathscr{O}(N \log m)$ time. Improving upon existing work on fuzzy data summarization, the new method uses a straightforward optimality criterion rather than heuristic quality indicators; the method is guaranteed to determine the optimal quantifier; and the optimal selection can be established very quickly. The resulting quantifier should be fitted to the quality of the data to improve the linguistic appropriateness of the summary. The relative proportion coefficients $r^{\min}$ and $r^{\max}$ on which the computation of the optimal quantifier is based, are interesting in their own right as a graphical representation of relative proportions found in imprecise data.

References

1. J. Barwise and R. Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219, 1981.
2. D. Dubois, H. Prade, and R.R. Yager, editors. *Fuzzy Information Engineering*. Wiley, New York, 1997.
3. I. Glöckner. DFS – an axiomatic approach to fuzzy quantification. TR97-06, Technical Faculty, University Bielefeld, 33501 Bielefeld, Germany, 1997.
4. Ingo Glöckner. *Fuzzy Quantifiers: A Computational Theory*, volume 193 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin, 2006.
5. I. Glöckner. Evaluation of quantified propositions in generalized models of fuzzy quantification. *International Journal of Approximate Reasoning*, 37(2):93–126, 2004.
6. J. Kacprzyk and P. Strykowski. Linguistic summaries of sales data at a computer retailer: A case study. In *Proc. IFSA '99*, pages 29–33, 1999.
7. D. Rasmussen and R.R. Yager. A fuzzy SQL summary language for data discovery. In Dubois et al. [2], pages 253–264.
8. R.R. Yager. A new approach to the summarization of data. *Information Sciences*, 28:69–86, 1982.
9. R.R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.
10. R.R. Yager. On linguistic summaries of data. In W. Frawley and G. Piatetsky-Shapiro, editors, *Knowledge Discovery in Databases*, pages 347–363. AAAI/MIT Press, 1991.
11. L.A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers and Mathematics with Applications*, 9:149–184, 1983.