Linguistic Summarization Using IF-THEN Rules

Dongrui Wu, *Member, IEEE*, Jerry M. Mendel, *Life Fellow, IEEE*, and Jhiin Joo, *Student Member, IEEE*

***Abstract*— Linguistic summarization (LS) is a data mining or knowledge discovery approach to extract patterns from databases. It has been studied by many researchers; however, none of them has used it to generate IF-THEN rules, which can be added to a knowledge base for better understanding of the data, or be used in Perceptual Reasoning to infer the outputs for new scenarios. In this paper LS using IF-THEN rules is proposed. Five quality measures for such summaries are defined. Among them, the *degree of usefulness* is especially valuable for finding the most reliable and representative rules, and the *degree of outlier* can be used to identify outlier rules and data. An example verifies the effectiveness of our approach. The relationship between LS and the Wang-Mendel method is also discussed.**

I. Introduction

The rapid progress of information technology has made huge amounts of data accessible to people. Unfortunately, the raw data alone are often hardly understandable and do not provide knowledge, i.e., frequently people face the "data rich, information poor" dilemma. Data mining approaches to automatically summarize the data and output human-friendly information are highly desirable. According to Mani and Maybury [19], "*summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).*" Particularly, data summarization in this paper means to [27] "*grasp and briefly describe trends and characteristics appearing in a dataset, without doing (explicit) manual 'record-by-record' analysis.*"

There can be two approaches to summarize a dataset: numerical summarization and linguistic summarization (LS). Statistical characteristics, such as mean, median, variance, etc, are examples of numerical summarization; however, as pointed out by Yager [43], "*summarization would be especially practicable if it could provide us with summaries that are not as terse as the mean, as well as treating the summarization of nonnumeric data.*" This suggests that LS of databases, which outputs summaries like "*About 1/2 of sales in autumn is of accessories*" [13]–[15] or "*IF X is large and Y is medium, THEN Z is small*," is more favorable, because it can provide richer and more easily understandable information, and it also copes well with nonnumeric data.

There are many approaches for LS of databases [4], [5], [32], [33] and time series [3], [12]. The fuzzy set (FS) based approach, introduced by Yager [43]–[46] and advanced by many others [6], [12], [15], [27], [34], is the most popular one. It can output summaries like "*About 1/2 of sales in autumn is of accessories*" and "*Most senior workers have high salary.*" Most of the works focus on type-1 (T1) FSs. Niewiadomski et al. [26]–[31] are to date the only ones working on LS using interval and general type-2 FSs [21], [47].

Dongrui Wu is with the Institute for Creative Technologies and the Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA 90089 (phone: 213-595-3269; email: dongruiw@usc.edu).

Jerry M. Mendel is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 (phone: 213-740-4445; email: mendel@sipi.usc.edu).

Jhiin Joo is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 (phone: 213-740-4456; email: jhijoo@usc.edu).

In this paper, we focus on LS using IF-THEN rules, e.g., "*IF X is large and Y is medium, THEN Z is small*," because our primary goal is to use LS to generate a rulebase for perceptual reasoning and decision-making [24], [25], [40], [41].

The rest of this paper is organized as follows: Section II introduces our LS approach to generate IF-THEN rules using T1 FSs and its associated quality measures. Section III extends the results in Section II to IT2 FSs. Section IV illustrates our LS approach by an example. Section V discusses the relationship between LS and the Wang-Mendel method. Section VI draws conclusions.

II. Linguistic Summarization Using T1 FSs

The main purpose of this paper is to propose a LS approach using IT2 FSs. For ease in understanding, we start with LS using T1 FSs; however, this does not mean we advocate that T1 FSs should be used in LS. In fact, we always argue that IT2 FSs should be used in LS, because they can model both intra-personal and inter-personal uncertainties [23], [25], as argued in Section III-A.

A. Data Description

For easy reference, our most frequently used symbols are summarized in Table I.

TABLE I
EXPLANATIONS OF THE SYMBOLS USED IN THIS PAPER. $n = 1, 2, \ldots, N$ AND $m = 1, 2, \ldots, M$.

	Meaning	Example
$\mathbb{D}$	The complete database	Haberman's Survival Dataset [1]
$\mathbb{Y}$	The set of all objects	All patients in the dataset
M	Number of objects in $\mathbb{Y}$	The total number of patients (306)
y_m	The m^{th} object	The m^{th} patient in the dataset
v_n	Name of the n^{th} attribute	Age (1^{st} attribute)
$\mathbb{X}_n$	The domain of v_n	[30, 83] for Age
$\mathbb{V}$	A set of all attribute names	<Age, Year, #Nodes, Survival>
v_n^m	Value of the n^{th} attribute for y_m	30 (Age of the 1^{st} patient)
$\mathbf{d}_m$	A complete record related to y_m	<30, 64, 1, Yes> for the 1^{st} patient
S_n	Summarizer	Around 35, very small #nodes
Q	Quantifier	Most, more than 100
w_g	Qualifier	Around 35, very small #nodes
T	Degree of truth	Any value in $[0, 1]$
C	Degree of sufficient coverage	Any value in $[0, 1]$
U	Degree of usefulness	Any value in $[0, 1]$
O	Degree of outlier	Any value in $[0, 1]$
S	Degree of simplicity	Any value in $[0, 1]$

Define a set of M objects $\mathbb{Y} = \{y_1, y_2, \ldots, y_M\}$ and a set of N attributes $\mathbb{V} = \{v_1, v_2, \ldots, v_N\}$. Let $\mathbb{X}_n$ $(n = 1, 2, \ldots, N)$ be the domain of v_n. Then, $v_n(y_m) \equiv v_n^m \in \mathbb{X}_n$ is the value of the n^{th} attribute for the m^{th} object $(m = 1, 2, \ldots, M)$. Hence, the database $\mathbb{D}$, which collects information about elements from $\mathbb{Y}$, is in the form of

$$\begin{aligned}\mathbb{D} &= \{< v_1^1, v_2^1, \ldots, v_N^1 >, \cdots, < v_1^M, v_2^M, \ldots, v_N^M >\} \\ &\equiv \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M\} \end{aligned} \quad (1)$$

where $\mathbf{d}_m =< v_1^m, v_2^m, \ldots, v_N^m >$ is a complete record about object y_m.

For example, for the Haberman's Survival Dataset [1] used in Section IV, there are 306 breast cancer patients ($M = 306$), and hence $\mathbb{Y}$={Patient1, Patient2, ..., Patient306}. Each

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

patient has four attributes ($N = 4$), and $\mathbb{V}$=<Age, Year, #Nodes, Survival>. For Age, its value ranges from 30 to 83; so, its domain $\mathbb{X}_1 = [30, 83]$. Patient1 was 30 years old, operated on in 1964, with one positive axillary node detected, and survived five years or longer. So, the complete record for Patient1 is $\mathbf{d}_1$=<30, 1964, 1, Yes>.

B. LS Using IF-THEN Rules and T1 FSs

Only single-antecedent single-consequent (SASC) rules are considered in this subsection. Multi-antecedent multi-consequent (MAMC) rules are considered in Section II-I.

Because we are interested in generating IF-THEN rules from a dataset, our *canonical form for LS using T1 FSs* is:

$$\text{IF } \mathbb{X}_1 \text{ is/has } S_1, \text{ THEN } \mathbb{X}_2 \text{ is/has } S_2 \quad [Q] \tag{2}$$

where S_1 and S_2 are words modeled by T1 FSs, and $Q \in [0, 1]$ is a *quality measure*, which indicates how good the rule is. One example of such a rule is:

$$\text{IF } \underbrace{\text{Age}}_{\mathbb{X}_1} \text{ is } \underbrace{\text{around 35}}_{S_1}, \text{ THEN } \underbrace{\text{survival}}_{\mathbb{X}_2} \text{ is } \underbrace{\text{yes}}_{S_2} \ [Q] \tag{3}$$

Once a dataset is given, the antecedents and consequents of the rules are determined. A user needs to specify the words used for each antecedent and consequent, and also their corresponding FS models. Then, all possible combinations of the rules can be constructed. The challenge is to compute Q, which can have different definitions.

C. Quality Measures of LS Using T1 FSs

According to Hirota and Pedrycz [8], the following five features are essential to measure the quality of a summary:

1) *Validity*: The summaries must be derived from data with high confidence.
2) *Generality*: This describes how many data support a summary.
3) *Usefulness*: This relates the summaries to the goals of the user, especially in terms of the impact that these summaries may have on decision-making.
4) *Novelty*: This describes the degree to which the summaries deviate from our expectations, i.e., how unexpected the summaries are.
5) *Simplicity*: This measure concerns the syntactic complexity of the summaries.

Next we propose five quality measures for T1 FS LS, corresponding to *validity*, *generality*, *usefulness*, *novelty* and *simplicity*, respectively.

D. Degree of Truth, T

Validity is represented by the *degree of truth*, T, which is computed as:

$$T = \frac{\sum_{m=1}^{M} \min(\mu_{S_1}(v_1^m), \mu_{S_2}(v_2^m))}{\sum_{m=1}^{M} \mu_{S_1}(v_1^m)} \tag{4}$$

Essentially, T is Kosko's subsethood measure [17] for T1 FSs and is also called *conditional and unqualified proposition* by Klir and Yuan [16]. This kind of formula has also been used in van den Berg et al.'s conditional probability for fuzzy events [35], and in computing the confidence of fuzzy association rules [9]–[11]. Roughly speaking, T increases as more data satisfying the antecedent also satisfy the consequent.

A different representation of the degree of truth T defined in (4) is introduced next. It will lead easily to the computation of T for LS using IT2 FSs, as will be shown in Section III-B. But first, two related definitions are introduced.

Definition 1: The *cardinality* of a T1 FS S_1 on database $\mathbb{D}$ is defined as

$$c_{\mathbb{D}}(S_1) = \sum_{m=1}^{M} \mu_{S_1}(v_1^m). \quad \square \tag{5}$$

Definition 2: The *joint cardinality* of T1 FSs $\{S_1, ..., S_N\}$ on database $\mathbb{D}$ is defined as

$$c_{\mathbb{D}}(S_1, ..., S_N) = \sum_{m=1}^{M} \min\{\mu_{S_1}(v_1^m), ..., \mu_{S_N}(v_N^m)\}. \quad \square \tag{6}$$

Using the cardinality $c_{\mathbb{D}}(S_1)$ and joint cardinality $c_{\mathbb{D}}(S_1, S_2)$, (4) can be re-expressed as:

$$T = \frac{c_{\mathbb{D}}(S_1, S_2)}{c_{\mathbb{D}}(S_1)}. \tag{7}$$

E. Degree of Sufficient Coverage, C

Generality is represented by the *degree of sufficient coverage*, C, which describes whether a rule is supported by enough data. To compute C, we first compute the *coverage ratio*, which is

$$r_c = \frac{\sum_{m=1}^{M} t_m}{M} \tag{8}$$

where

$$t_m = \begin{cases} 1, & \mu_{S_1}(v_1^m) > 0 \text{ and } \mu_{S_2}(v_2^m) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

i.e., r_c is the percentage of data which fit both the antecedent and the consequent of the rule. Because each rule only covers a small region of the high-dimensional input-output space, r_c is usually very small (e.g., mostly smaller than 0.1). So, $r_c = 0.15$ may be considered sufficient coverage with degree 1. The following mapping converts the coverage ratio into the appropriate degree of sufficient coverage, and agrees with our feeling about sufficient coverage:

$$C = f(r_c) \tag{10}$$

where f is a function that maps r_c into C. The S-shape function $f(r_c)$ used in this paper is shown in Fig. 1. It is determined by two parameters r_1 and r_2 ($0 \le r_1 < r_2$), i.e.,

$$f(r_c) = \begin{cases} 0, & r_c \le r_1 \\ 2\left(\frac{r_c - r_1}{r_2 - r_1}\right)^2, & r_1 < r_c < \frac{r_1 + r_2}{2} \\ 1 - 2\left(\frac{r_2 - r_c}{r_2 - r_1}\right)^2, & \frac{r_1 + r_2}{2} \le r_c < r_2 \\ 1, & r_c \ge r_2 \end{cases} \tag{11}$$

and $r_1 = 0.02$ and $r_2 = 0.15$ are used in this paper. $f(r_c)$ can be modified according to the user's requirement about sufficient coverage.

F. Degree of Usefulness, U

The *degree of usefulness*, U, as its name suggests, describes how useful a summary is. A rule is useful if and only if:

1) It has high degree of truth, i.e., most of the data satisfying the rule's antecedents also have the behavior described by its consequent.
2) It has sufficient coverage, i.e., enough data are described by it.

Hence, U is computed as

$$U = \min(T, C) \tag{12}$$

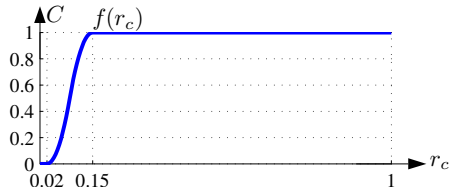

Fig. 1. The S-shape function $f(r_c)$ used in this paper.

G. Degree of Outlier, O

Novelty means *unexpectedness*. In this paper unexpectedness is represented by the *degree of outlier*, O, which indicates the possibility that a rule describes only outliers instead of a useful pattern. Clearly, the degree of sufficient coverage for an outlier rule must be very small, i.e., it only describes very few data; however, small C alone is not enough to identify outliers rules, and the degree of truth should also be considered. When C is small, T can be small (close to 0), medium (around 0.5) or large (close to 1), as shown in Fig. 2, where the rule "IF v_1 is Low, THEN v_2 is High" is illustrated for three different cases:

1) For the rule illustrated by the shaded region in Fig. 2(a), T is large because all data satisfying the antecedent (v_1 is Low) also satisfy the consequent (v_2 is High), i.e., $\sum_{m=1}^{M} \min(\mu_{Low}(v_1^m), \mu_{High}(v_2^m))$ is close to $\sum_{m=1}^{M} \mu_{Low}(v_1^m)$. Visual inspection suggests that this rule should be considered as an outlier because the data described by it are isolated from the rest.
2) For the rule illustrated by the shaded region in Fig. 2(b), T is small because most data satisfying the antecedent (v_1 is Low) do not satisfy the consequent (v_2 is High), i.e., $\sum_{m=1}^{M} \min(\mu_{Low}(v_1^m), \mu_{High}(v_2^m))$ much smaller than $\sum_{m=1}^{M} \mu_{Low}(v_1^m)$. Visual inspection suggests that this rule should also be considered as an outlier because the data described by it are isolated from the rest.
3) For the rule illustrated by the shaded region in Fig. 2(c), T is medium because the data satisfying the antecedent (v_1 is Low) are distributed somewhat uniformly in the v_2 domain, i.e., $\sum_{m=1}^{M} \min(\mu_{Low}(v_1^m), \mu_{High}(v_2^m))$ is about half of $\sum_{m=1}^{M} \mu_{Low}(v_1^m)$. By visual inspection, this rule should not be considered as an outlier (although it is not a good rule as U would be small) because its data are not so isolated from the rest.

In summary, an outlier rule must satisfy:

1) The degree of truth, T, must be very small or very large.
2) The degree of sufficient coverage, C, must be very small.

Finally, note that the purpose of finding an outlier rule is to help people identify possible outlier data and then to further investigate them. So, we need to exclude a rule with $T = 0$ from being identified as an outlier because in this case the rule does not describe any data. The following formula is used in this paper to compute the degree of outlier:

$$O = \begin{cases} \min(\max(T, 1-T), 1-C), & T > 0 \\ 0, & T = 0 \end{cases} \quad (13)$$

The term $\max(T, 1-T)$ converts a small T (close to 0) or a large T (close to 1) to a large number in [0, 1],

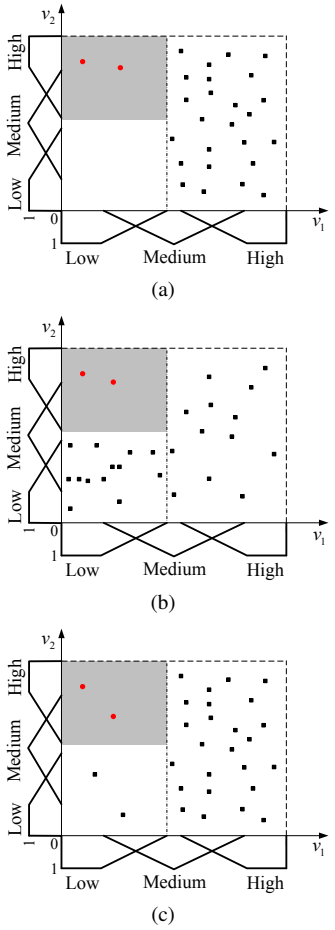

Fig. 2. Three cases for the rule "IF v_1 is Low, THEN v_2 is High," whose C is small. (a) T is large, (b) T is small, and (c) T is medium.

which is required by the first criterion of an outlier rule, and $\min(\max(T, 1-T), 1-C)$ further imposes the constraint that C must be small, which is the second criterion for an outlier rule. Note that the closer O is to 1, the more a rule is judged to be an outlier.

A graph illustrating the dependence of U in (12) and O in (13) on T and C is shown in Fig. 3. U or O increases as (T, C) moves in the directions indicated by the arrows, e.g., U moves toward 1 as both T and C increase.

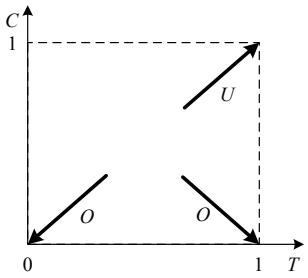

Fig. 3. Illustration of useful rules and outlier rules determined by T and C.

H. Degree of Simplicity, S

The *simplicity* of a summary can be measured by its length, i.e., how many antecedents and consequents the rule has. We

define the *degree of simplicity*, S, of a rule by:

$$S = 2^{2-l} \tag{14}$$

where l is the total number of antecedents and consequents of the rule. Clearly, $S \in (0, 1]$, and the simplest rule ($S = 1$) has only one antecedent and one consequent. As the number of antecedents and/or consequents increases, S decreases, and a rule becomes more difficult to understand.

I. Multi-Antecedent Multi-Consequent (MAMC) Rules

The generalization of the results for SASC rules to MAMC rules is straightforward. Consider an MAMC rule:

$$\text{IF } \mathbb{X}_1 \text{ is/has } S_1 \text{ and ... and } \mathbb{X}_K \text{ is/has } S_K, \text{ THEN } \mathbb{X}_{K+1} \text{ is/has } S_{K+1} \text{ and ... and } \mathbb{X}_N \text{ is/has } S_N \quad [Q] \tag{15}$$

The degree of truth, T, is computed as

$$T = \frac{c_{\mathbb{D}}(S_1, ..., S_N)}{c_{\mathbb{D}}(S_1, ..., S_K)} \tag{16}$$

and C is computed by redefining t_m as

$$t_m = \begin{cases} 1, & \mu_{S_n}(v_n^m) > 0, \quad \forall n = 1, ..., N \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Once r_c is obtained, C is computed by (10). Because both T and C are crisp numbers, (12) and (13) can again be used to compute U and O. S is still computed by (14).

III. Linguistic Summarization Using IT2 FSs

In this section we extend the results in the previous section to IT2 FSs.

A. Why IT2 FSs Should Be Used to Model Words

People communicate using words. There are at least two types of uncertainties associated with a word [22], [37]: *intra-personal uncertainty* and *inter-personal uncertainty*. Intra-personal uncertainty describes [22] "*the uncertainty a person has about the word*." It is also explicitly pointed out by Wallsten and Budescu [37] as *"except in very special cases, all representations are vague to some degree in the minds of the originators and in the minds of the receivers,"* and they suggest to model it by a T1 FS. Inter-personal uncertainty describes [22] "*the uncertainty that a group of people have about the word*." It is pointed out by Mendel [21] as *"words mean different things to different people"* and Wallsten and Budescu [37] as *"different individuals use diverse expressions to describe identical situations and understand the same phrases differently when hearing or reading them."* Because an IT2 FS has an FOU which can be viewed as a group of T1 FSs (see Fig. 4), it can model both types of uncertainty [22]; hence, we suggest IT2 FSs be used in modeling words [20]–[22], [25], [40]. Additionally, Mendel [23] has explained why it is scientifically incorrect to model a word using a T1 FS, i.e., (1) A T1 FS for a word is well-defined by its membership function (MF) that is totally certain once all of its parameters are specified; (2) words mean different things to different people, and so are uncertain; and, therefore, (3) it is a contradiction to say that something certain can model something that is uncertain.

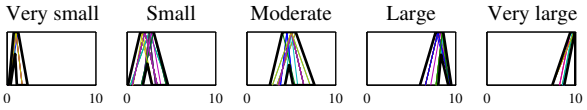

Fig. 4. Five examples of word FOUs obtained from the Interval Approach [18]. The areas between the thick curves are FOUs, and the curves within the FOUs are embedded T1 FSs mapped from individuals' endpoint data.

B. LS Using IF-THEN Rules and IT2 FSs

When IT2 FSs are used in a LS to generate IF-THEN rules, our canonical form in (2) becomes:

$$\text{IF } \mathbb{X}_1 \text{ is/has } \tilde{S}_1, \text{ THEN } \mathbb{X}_2 \text{ is/has } \tilde{S}_2 \quad [Q] \tag{18}$$

where $\tilde{S}_1$ and $\tilde{S}_2$ are words modeled by IT2 FSs, and $Q \in [0, 1]$ is a quality measure.

Next we explain how to compute the five different quality measures.

C. Quality Measures for LS Using IT2 FSs

Recall from (7) that the degree of truth for LS using T1 FSs is computed based on the cardinalities of T1 FSs on a database $\mathbb{D}$. To extend that result to IT2 FSs, the following definitions are needed.

Definition 3: The *cardinality* of an IT2 FS $\tilde{S}_1$ on dataset $\mathbb{D}$ is defined as

$$C_{\mathbb{D}}(\tilde{S}_1) \equiv [c_{\mathbb{D}}(\underline{S}_1), c_{\mathbb{D}}(\overline{S}_1)] = \left[\sum_{m=1}^{M} \mu_{\underline{S}_1}(v_1^m), \sum_{m=1}^{M} \mu_{\overline{S}_1}(v_1^m)\right] \tag{19}$$

and the *average cardinality* is

$$c_{\mathbb{D}}(\tilde{S}_1) = \frac{c_{\mathbb{D}}(\underline{S}_1) + c_{\mathbb{D}}(\overline{S}_1)}{2}. \quad \square \tag{20}$$

Definition 4: The *joint cardinality* of IT2 FSs $\{\tilde{S}_1, ..., \tilde{S}_N\}$ on database $\mathbb{D}$ is defined as

$$\begin{aligned} C_{\mathbb{D}}(\tilde{S}_1, ..., \tilde{S}_N) &\equiv \left[c_{\mathbb{D}}(\underline{S}_1, ..., \underline{S}_N), c_{\mathbb{D}}(\overline{S}_1, ..., \overline{S}_N)\right] \\ &= \left[\sum_{m=1}^{M} \min\{\mu_{\underline{S}_1}(v_1^m), ..., \mu_{\underline{S}_N}(v_N^m)\}, \right. \\ &\quad \left. \sum_{m=1}^{M} \min\{\mu_{\overline{S}_1}(v_1^m), ..., \mu_{\overline{S}_N}(v_N^m)\}\right] \end{aligned} \tag{21}$$

and the *average joint cardinality* is

$$c_{\mathbb{D}}(\tilde{S}_1, ..., \tilde{S}_N) = \frac{c_{\mathbb{D}}(\underline{S}_1, ..., \underline{S}_N) + c_{\mathbb{D}}(\overline{S}_1, ..., \overline{S}_N)}{2}. \quad \square \tag{22}$$

By substituting the cardinalities in (7) by their respective average cardinalities, T in (18) is computed as

$$T = \frac{c_{\mathbb{D}}(\tilde{S}_1, \tilde{S}_2)}{c_{\mathbb{D}}(\tilde{S}_1)}. \tag{23}$$

which is essentially Vlachos and Sergiadis's subsethood measure [36], [40], [42] for interval-valued fuzzy sets.

For LS using IT2 FSs, r_c is still computed by (8), but t_m is defined differently:

$$t_m = \begin{cases} 1, & \mu_{\overline{S}_1}(v_1^m) > 0 \text{ and } \mu_{\overline{S}_2}(v_2^m) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

i.e., we count all objects with non-zero membership on both antecedent and consequent. Once r_c is obtained, C is computed by (10).

Because both T and C are crisp numbers, (12) and (13) can again be used to compute U and O. S is still computed by (14).

D. Multi-Antecedent Multi-Consequent Rules

The generalization of the results for SASC rules to MAMC rules is straightforward. Consider an MAMC rule:

$$\text{IF } \mathbb{X}_1 \text{ is/has } \tilde{S}_1 \text{ and ... and } \mathbb{X}_K \text{ is/has } \tilde{S}_K, \text{ THEN}$$
$$\mathbb{X}_{K+1} \text{ is/has } \tilde{S}_{K+1} \text{ and ... and } \mathbb{X}_N \text{ is/has } \tilde{S}_N \quad [Q] \quad (25)$$

The degree of truth, T, is computed as

$$T = \frac{c_{\mathbb{D}}(\tilde{S}_1, ..., \tilde{S}_N)}{c_{\mathbb{D}}(\tilde{S}_1, ..., \tilde{S}_K)} \quad (26)$$

and r_c is computed by redefining t_m as

$$t_m = \begin{cases} 1, & \mu_{\overline{S}_n}(v_n^m) > 0, \quad \forall n = 1, ..., N \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

Once r_c is obtained, C is computed by (10). Because both T and C are crisp numbers, (12) and (13) can again be used to compute U and O. S is still computed by (14).

IV. Example

The Haberman's survival dataset [1] is used as an example to illustrate our LS approach. It contains 306 cases on the survival of patients who had undergone surgery for breast cancer. LS was used to find the relationship between the following inputs and *whether or not a patient survived 5 years or longer*:

1) *Age:* The age of the patient at the time of operation.
2) *Year:* The patient's year of operation.
3) *#Nodes:* The number of positive axillary nodes detected.

Figs. 5-8 show the top 10 rules when T, C, U and O are used as the ranking criterion, respectively. The cases are displayed by a Parallel Coordinates approach [2] in the middle of the GUI, where each coordinate represents an attribute, and the two numbers labeled at the two ends of each coordinate represent the range of that attribute, e.g., observe from Fig. 5 that *Age* has range [30, 83]. Each case is represented in the middle of Fig. 5 as a piece-wise linear curve. The blue curves represent those cases supporting the current rule under consideration (i.e., those cases satisfying *both* the antecedents and the consequent of the rule), and the strength of supporting is proportional to the depth of the blue color. The red curves represent those cases violating the current rule (i.e., those cases satisfying *only* the antecedents of the rule), and the strength of violating is proportional to the depth of the red color. The black curves are cases irrelevant to the current rule (i.e., those cases *not* satisfying the antecedents of the rule). The light green region indicates the area covered by the current rule.

The bottom axes in Fig. 5 shows the IT2 FSs used for each attribute. The IT2 FSs that are used in the current rule are highlighted in green and their names are also displayed.

Observe:

1) From Fig. 5, when T is used as the ranking criterion, a rule with high T may describe only very few cases, so it is very possible that this rule only describes outliers and hence cannot be trusted. This suggests that T alone is not a reliable quality measure for LS.
2) From Fig. 6, when C is used as the ranking criterion, a rule with high C may have a low degree of truth. So, C alone is not a good quality measure, either.
3) From Fig. 7, when U is used as the ranking criterion, a rule with high U has both high degree of truth and sufficient coverage, and hence it describes a useful rule. So, U is a comprehensive and reliable quality measure for LS.
4) From Fig. 8, when O is used as the ranking criterion, a rule with high O usually describes a very small number of cases, which should be considered as outliers. So, O is useful in finding unexpected data and rules.

In summary, it appears that U and O are better quality measures for LS than T which is dominant in previous LS literature: a high U identifies a useful rule with both high degree of truth and sufficient coverage, whereas a high O identifies outliers in the dataset that are worthy of further investigation.

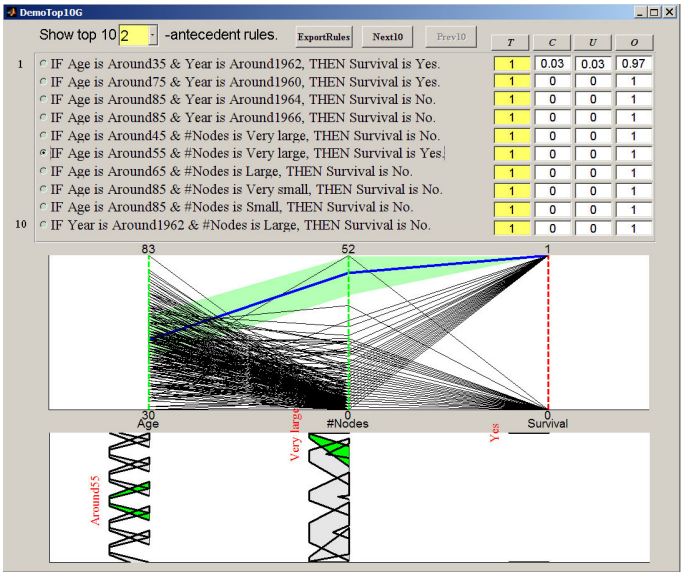

Fig. 5. Haberman's Survival Dataset: Top 10 rules according to T, the degree of truth. The middle and bottom parts illustrate the 6^{th} rule.

Finally, note that Figs. 5-8 are only used to illustrate the difference among the four quality measures. In practice, linguistic summarization can be made more useful by specifying an (or several) antecedent and study what rules it leads to, e.g., what are the top rules if *Age is Around35*, or by specifying a consequent and study what combinations of antecedents lead to it, e.g., what antecedents lead to the consequent *Survival is No*. Due to space limit, the details will be reported in a forthcoming journal article.

V. Discussions

In this section the relationship between LS and the Wang-Mendel (WM) method [21], [39], a simple yet effective method to generate fuzzy rules from training examples, is discussed. Because currently the WM method mainly focuses on T1 FSs, only T1 FSs are used in the discussion; however, our results can be extended to IT2 FSs without problems.

We use Fig. 9, where the 18 training data points are represented by squares[1], to introduce its idea:

1) Each input (x) and output (y) domain is partitioned into $2L + 1$ (an odd number) overlapping intervals, where L can be different for each variable. Then, MFs

[1]Three points are represented by different shapes only for easy reference purpose.

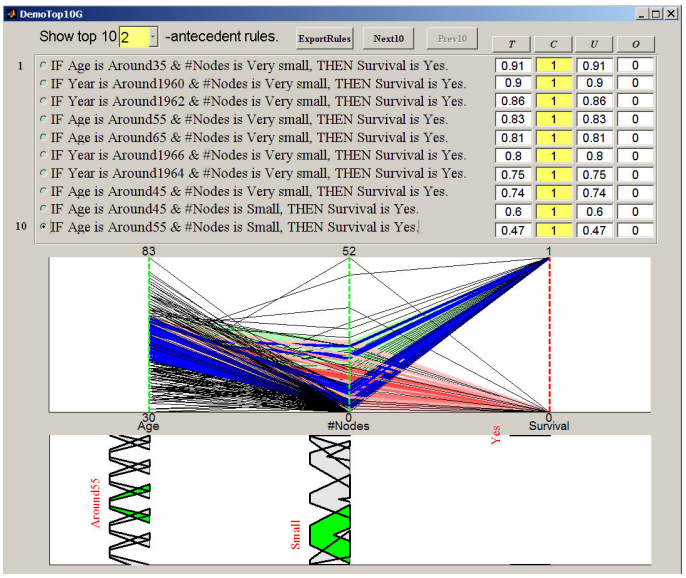

Fig. 6. Haberman's Survival Dataset: Top 10 rules according to C, the degree of sufficient coverage. The middle and bottom parts illustrate the 10^{th} rule.

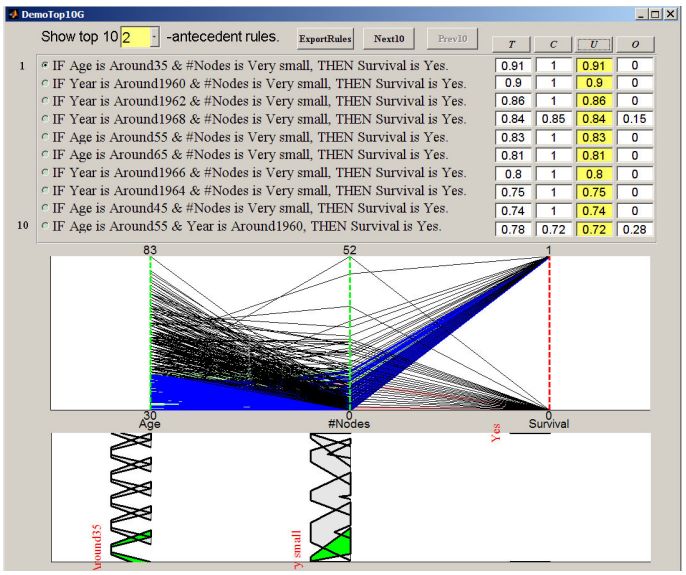

Fig. 7. Haberman's Survival Dataset: Top 10 rules according to U, the degree of usefulness. The middle and bottom parts illustrate the first rule.

and labels are assigned to these intervals. In Fig. 9, each of the x and y domains is partitioned into three overlapping intervals by the FSs Low, Medium and High. An interval in the x domain and an interval in the y domain together determine a region in the input-output space, e.g., the region determined by High x and Low y is shown as the shaded region in the lower right corner of Fig. 9.

2) Because of overlapping MFs, it frequently happens that a datum is in more than one region, e.g., the diamond in Fig. 9 belongs to the region determined by High x and Low y, and also to the region determined by High x and Medium y. For each (x, y), one evaluates its degrees of belonging in regions where it occurs, assigns it to the region with maximum degree, and generates a rule from it. For example, the degree of belonging of the diamond in Fig. 9 to the region determined by High x and Low y (the shaded region in the lower right corner) is $\mu_{High}(x)\mu_{Low}(y) = 1 \times 0.1 = 0.1$, and its degree of belonging to the region determined by High x and Medium y is $\mu_{High}(x)\mu_{Medium}(y) = 1 \times 0.8 = 0.8$; so, the diamond should be assigned to the region determined by High x and Medium y. Consequently, the corresponding rule generated from this diamond is

$$\text{IF } x \text{ is High, THEN } y \text{ is Medium} \tag{28}$$

and it is also assigned a degree of 0.8. Similarly, a rule generated from the cross in Fig. 9 is

$$\text{IF } x \text{ is High, THEN } y \text{ is Low} \tag{29}$$

and it has a degree of $\mu_{High}(x)\mu_{Low}(y) = 1 \times 1 = 1$.

3) To resolve conflicting rules, i.e., rules with the same antecedent MFs and different consequent MFs, one chooses the rule with the highest degree and discards all other rules, e.g., Rules (28) and (29) are conflicting, and Rule (29) is chosen because it has a higher degree.

Finally, the three rules generated by the WM method for the Fig. 9 data are:

$$\text{IF } x \text{ is Low, THEN } y \text{ is High}$$
$$\text{IF } x \text{ is Medium, THEN } y \text{ is Medium}$$
$$\text{IF } x \text{ is High, THEN } y \text{ is Low}$$

The first rule seems counter-intuitive, but it is a true output of the WM method. It is generated by the circle in Fig. 9 with a degree $\mu_{Low}(x)\mu_{High}(y) = 1 \times 1 = 1$, i.e., its degree is higher than two other possible rules, IF x is Low, THEN y is Low and IF x is Low, THEN y is Medium, through these two rules have more data to support them and hence look more reasonable. Note, however, that this example considers an extreme case. In practice the WM method usually generates very reasonable rules, which is why it is popular.

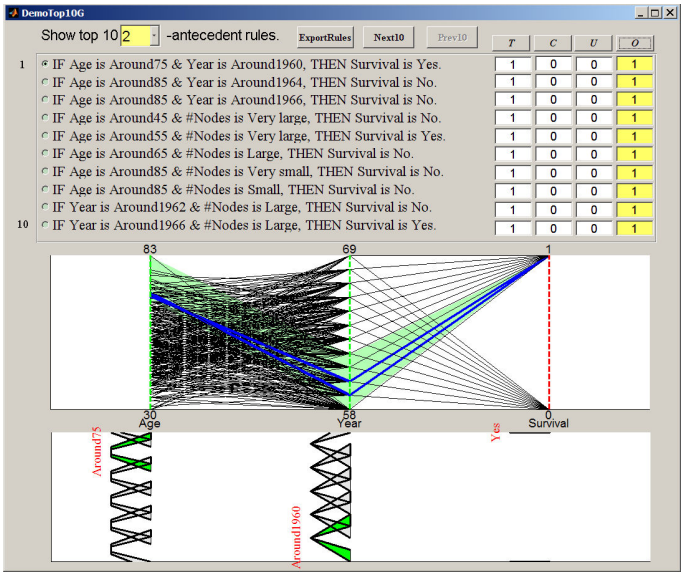

Fig. 8. Haberman's Survival Dataset: Top 10 rules according to O, the degree of outlier. The middle and bottom parts illustrate the first rule.

Once the rules are generated, the degrees associated with them are discarded as they are no longer useful.

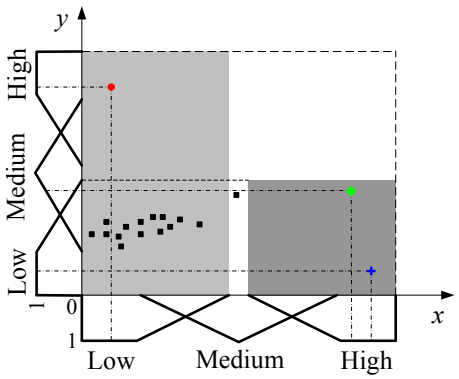

Fig. 9. An example to illustrate the difference between the WM method and LS. When x is Low, the WM method generates a rule "IF x is Low, THEN y is High" whereas LS generates a rule "IF x is Low, THEN y is Low."

Example 1: Fig. 9 can also be used to illustrate the difference between the WM method and LS. Consider the shaded region where x is Low. There are three candidates for a rule in this region:

$$\text{IF } x \text{ is Low, THEN } y \text{ is High} \tag{30}$$
$$\text{IF } x \text{ is Low, THEN } y \text{ is Medium} \tag{31}$$
$$\text{IF } x \text{ is Low, THEN } y \text{ is Low} \tag{32}$$

For Rule (30),

$$c_{\mathbb{D}}(\text{Low}_x, \text{High}_y) = \sum_{m=1}^{18} \min(\mu_{\text{Low}_x}(x_m), \mu_{\text{High}_y}(y_m)) = 1 \tag{33}$$

$$c_{\mathbb{D}}(\text{Low}_x) = \sum_{m=1}^{18} \mu_{\text{Low}_x}(x_m) = 12.8 \tag{34}$$

$$T = \frac{c_{\mathbb{D}}(\text{Low}_x,\ \text{High}_y)}{c_{\mathbb{D}}(\text{Low}_x)} = 0.08 \tag{35}$$

Because the dataset consists of 18 points and there is only one datum that falls in the region determined by Low x and High y, the coverage ratio [see (8)] and degree of sufficient coverage [see (10)] are

$$r_c = 1/18 \tag{36}$$
$$C = f(r_c) = 0.15 \tag{37}$$

and hence $U = \min(T, C) = 0.08$ and $O = \min(\max(T, 1-T), 1-C) = \min(\max(0.08, 0.92), 1-0.15) = 0.85$.

Similarly, for Rule (31) LS gives:

$$T = 0.31, \quad C = 1, \quad U = 0.31, \quad O = 0 \tag{38}$$

and for Rule (32), LS gives:

$$T = 0.71, \quad C = 1, \quad U = 0.71, \quad O = 0 \tag{39}$$

By ranking U and O, LS would select Rule (32) as the most useful rule with $U = 0.71$ and Rule (30) as an outlier with $O = 0.85$. These results are more reasonable than the rules generated by the WM method.

Repeating the above procedure for the other two regions, the following three rules are generated when U is used as the ranking criterion:

IF x is Low, THEN y is Low
$T = 0.71,\ C = 1,\ U = 0.71,\ O = 0$
IF x is Medium, THEN y is Medium
$T = 0.82,\ C = 1,\ U = 0.82,\ O = 0$
IF x is High, THEN y is Low
$T = 0.57,\ C = 0.82,\ U = 0.57,\ O = 0.18.$ □

In summary, the differences between the WM method and LS are:

1) The WM method tries to construct a predictive model whereas LS tries to construct a descriptive model. According to [7], "*a* descriptive model *presents, in convenient form, the main features of the data. It is essentially a summary of the data, permitting us to study the most important aspects of the data without their being obscured by the sheer size of the data set. In contrast, a* predictive model *has the specific objective of allowing us to predict the value of some target characteristic of an object on the basis of observed values of other characteristics of the object.*"
2) Both methods partition the problem domain into several smaller regions and try to generate a rule for each region; however, the WM method generates a rule for a region as long as there are data in it, no matter how many data are there, whereas LS does not, e.g., if a region has very few data in it, then these data may be considered as outliers and no useful rule is generated for this region.
3) The rules obtained from LS have several quality measures associated with them, so the rules can be sorted according to different criteria, whereas the rules obtained from the WM method are considered equally important[2].

VI. Conclusions

LS is a data mining or knowledge discovery approach to extract patterns from databases. Many authors have used this technique to generate summaries like "Most senior workers have high salary," which can be used to better understand and communicate about data; however, none of them has used it to generate IF-THEN rules like "IF X is large and Y is medium, THEN Z is small," which not only facilitate understanding and communication of data, but also can be used in decision-making. In this paper a LS approach to generate IF-THEN rules has been proposed. Both type-1 and interval type-2 fuzzy sets are considered. Five quality measures for such summaries have been proposed:

1) The degree of truth, which quantifies the validity of a rule.
2) The degree of sufficient coverage, which describes how many data support a rule and is related to the generality of the rule.
3) The degree of usefulness, which finds rules with both high validity and sufficient coverage.
4) The degree of outlier, which describes the novelty of a rules, i.e., the degree to which the summaries deviate from our expectations.
5) The degree of simplicity, which quantifies the syntactic complexity of the summaries.

[2]There is an improved version of the WM method [38] that assigns a degree of truth to each rule; however, the degree of truth is computed differently from T in this paper, and the rule consequents are numbers instead of words modeled by FSs; so, it is not considered in this paper.

Among them, the degree of usefulness is especially useful in finding the most reliable and representative rules, and the degree of outlier can be used to identify outlier rules and data for close-up investigation. These five quality measures also correspond to the concepts of validity, generality, usefulness, novelty and simplicity, five essential measures of a summary proposed by Hirota and Pedrycz [8].

Our future work includes:

1) To further study the applications of LS, e.g., how to use LS to rank the importance of inputs and hence to select the most important ones.
2) To design more efficient algorithms for LS. Currently we use an exhaustive search method, where all possible combinations of rules are evaluated and then ranked according to a certain quality measure to find the top rules. The computational cost of this approach increases rapidly when the size of the database increases, and/or the number of antecedents increases, and/or the number of FSs associated with each attribute increases. More efficient algorithms are necessary to facilitate the applications of LS. One idea is to use some heuristics to eliminate some less promising rules from evaluation.

ACKNOWLEDGEMENT

This study was funded by the Center of Excellence for Research and Academic Training on Interactive Smart Oilfield Technologies (CiSoft); CiSoft is a joint University of Southern California–Chevron initiative.

REFERENCES

[1] "Haberman's survival data set." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival.
[2] "Xmdv tool home page." [Online]. Available: http://davis.wpi.edu/~xmdv/.
[3] D. A. Chiang, L. R. Chow, and Y. F. Wang, "Mining time series data by a fuzzy linguistic summary system," *Fuzzy Sets and Systems*, vol. 112, pp. 419–432, 2000.
[4] D. Dubois and H. Prade, "Gradual rules in approximate reasoning," *Information Sciences*, vol. 61, pp. 103–122, 1992.
[5] W. Duch, R. Setiono, and J. Zurada, "Computational intelligence methods for rule-based data understanding," *Proc. IEEE*, vol. 92, no. 5, pp. 771–805, 2004.
[6] R. George and R. Srikanth, "Data summarization using genetic algorithms and fuzzy logic," in *Genetic Algorithms Soft Comput.*, F. Herrera and J. Verdegay, Eds. Heidelberg, Germany: Springer-Verlag, 1996, pp. 599–611.
[7] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Boston, MA: MIT Press, 2001.
[8] K. Hirota and W. Pedrycz, "Fuzzy computing for data mining," *Proc. IEEE*, vol. 87, no. 9, pp. 1575–1600, 1999.
[9] T. P. Hong, C. S. Kuo, and S. C. Chi, "Trade-off between computation time and number of rules for fuzzy mining from quantitative data," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 5, pp. 587–604, 2001.
[10] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," *Information Sciences*, vol. 136, no. 1-4, pp. 109–133, 2001.
[11] H. Ishibuchi and T. Yamamoto, "Rule weight specification in fuzzy rule-based classification systems," *IEEE Trans. on Fuzzy Systems*, vol. 13, no. 4, pp. 428–435, 2005.
[12] J. Kacprzyk, A. Wilbik, and S. Zadrozny, "Linguistic summarization of time series using a fuzzy quantifier driven aggregation," *Fuzzy Sets and Systems*, vol. 159, pp. 1485–1499, 2008.
[13] J. Kacprzyk and R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of of General Systems*, vol. 30, pp. 133–154, 2001.
[14] J. Kacprzyk, R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of of Applied Mathematics and Computer Science*, vol. 10, pp. 813–834, 2000.
[15] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: Towards natural language based knowledge discovery tools," *Information Sciences*, vol. 173, pp. 281–304, 2005.
[16] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ: Prentice-Hall, 1995.
[17] B. Kosko, "Fuzziness vs. probability," *International Journal of General Systems*, vol. 17, pp. 211–240, 1990.
[18] F. Liu and J. M. Mendel, "Encoding words into interval type-2 fuzzy sets using an interval approach," *IEEE Trans. on Fuzzy Systems*, vol. 16, no. 6, pp. 1503–1521, 2008.
[19] I. Mani and M. Maybury, *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1989.
[20] J. M. Mendel, "Computing with words, when words can mean different things to different people," in *Proc. 3rd Int'l ICSC Symp. on Fuzzy Logic and Applications*, Rochester, NY, June 1999, pp. 158–164.
[21] ——, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Upper Saddle River, NJ: Prentice-Hall, 2001.
[22] ——, "Computing with words and its relationships with fuzzistics," *Information Sciences*, vol. 177, pp. 988–1006, 2007.
[23] ——, "Computing with words: Zadeh, Turing, Popper and Occam," *IEEE Computational Intelligence Magazine*, vol. 2, pp. 10–17, 2007.
[24] J. M. Mendel and D. Wu, "Perceptual reasoning for perceptual computing," *IEEE Trans. on Fuzzy Systems*, vol. 16, no. 6, pp. 1550–1564, 2008.
[25] ——, *Perceptual Computing: Aiding People in Making Subjective Judgments*. Hoboken, NJ: Wiley-IEEE Press, 2010.
[26] A. Niewiadomski, *Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions*. Portland: Warszawa, 2008.
[27] ——, "A type-2 fuzzy approach to linguistic summarization of data," *IEEE Trans. on Fuzzy Systems*, vol. 16, no. 1, pp. 198–212, 2008.
[28] A. Niewiadomski and M. Bartyzel, "Elements of type-2 semantics in summarizing databases," *Lecture Notes in Artificial Intelligence*, vol. 4029, pp. 278–287, 2006.
[29] A. Niewiadomski and P. Szczepaniak, "News generating based on type-2 linguistic summaries of databases," in *Proc. Int'l Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Paris, France, July 2006, pp. 1324–1331.
[30] A. Niewiadomski, "On two possible roles of type-2 fuzzy sets in linguistic summaries," *Lecture Notes in Computer Science*, vol. 3528, pp. 341–347, 2005.
[31] ——, "Type-2 fuzzy summarization of data: An improved news generating," *Lecture Notes in Computer Science*, vol. 4585, pp. 241–250, 2007.
[32] G. Raschia and N. Mouaddib, "Using fuzzy labels as background knowledge for linguistic summarization of databases," in *Proc. IEEE Int'l Conf. on Fuzzy Systems*, Melbourne, Australia, December 2001, pp. 1372–1375.
[33] D. Rasmussen and R. Yager, "Finding fuzzy and gradual functional dependencies with SummarySQL," *Fuzzy Sets and Systems*, vol. 106, pp. 131–142, 1999.
[34] R. Saint-Paul, G. Raschia, and N. Mouaddib, "Database summarization: The SaintEtiQ system," in *Proc. IEEE Int'l Conf. on Data Engineering*, Istanbul, Turkey, April 2007, pp. 1475–1476.
[35] J. van den Berg, U. Kaymak, and W.-M. van den Bergh, "Fuzzy classification using probability-based rule weighting," in *Proc. IEEE Int'l Conf. on Fuzzy Systems*, Honolulu, Hawaii, May 2002, pp. 991–996.
[36] I. Vlachos and G. Sergiadis, "Subsethood, entropy, and cardinality for interval-valued fuzzy sets – an algebraic derivation," *Fuzzy Sets and Systems*, vol. 158, pp. 1384–1396, 2007.
[37] T. S. Wallsten and D. V. Budescu, "A review of human linguistic probability processing: General principles and empirical evidence," *The Knowledge Engineering Review*, vol. 10, no. 1, pp. 43–62, 1995.
[38] H. Wang and D. Qiu, "Computing with words via Turing machines: A formal approach," *IEEE Trans. on Fuzzy Systems*, vol. 11, no. 6, pp. 742–753, 2003.
[39] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 22, no. 2, pp. 1414–1427, 1992.
[40] D. Wu, "Intelligent systems for decision support," Ph.D. dissertation, University of Southern California, Los Angeles, CA, May 2009.
[41] D. Wu and J. M. Mendel, "Perceptual reasoning for perceptual computing: A similarity-based approach," *IEEE Trans. on Fuzzy Systems*, vol. 17, no. 6, pp. 1397–1411, 2009.
[42] ——, "Interval type-2 fuzzy set subsethood measures as a decoder for perceptual reasoning," Signal and Image Processing Institute, University of Southern California, Los Angeles, CA, Tech. Rep. USC-SIPI Report 398, 2010.
[43] R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, pp. 69–86, 1982.
[44] ——, "On linguistic summaries of data," in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and B. Frawley, Eds. MIT Press, 1991, pp. 347–363.
[45] ——, "Linguistic summaries as a tool for database discovery," in *Proc. IEEE Int'l Conf. on Fuzzy Systems*, Yokohama, Japan, 1995, pp. 79–82.
[46] ——, "Database discovery using fuzzy sets," *International Journal of Intelligent Systems*, vol. 11, pp. 691–712, 1996.
[47] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-1," *Information Sciences*, vol. 8, pp. 199–249, 1975.