Generating Quality IF-THEN Rules for Diabetes using Linguistic Summarization

Priyanka Mehra
Dept. of Computer Science
South Asian University,
Chanakyapuri,
New Delhi 110021, India
Email: pmehra98.pm@gmail.com

Taniya Seth
Dept. of Computer Science
South Asian University,
Chanakyapuri,
New Delhi 110021, India
Email: taniya.seth@students.sau.ac.in

Pranab K. Muhuri
Dept. of Computer Science
South Asian University,
Chanakyapuri,
New Delhi 110021, India
Email: pranabmuhuri@cs.sau.ac.in

Abstract—**Linguistic summarization is an approach of extraction of knowledge or linguistic patterns from datasets. Since summaries are brief, they promote quick analysis of data. Numerous researchers have utilized this approach to produce summaries which are easy to comprehend. Furthermore, linguistic summarization has been used to generate IF-THEN rules in the literature which not only convey data easily but is utilized in making decisions. In this paper we follow this approach to generate IF-THEN rules for diabetes on a dataset. This constructed dataset consists of responses from individuals for five parameters, crucial in the diagnosis of diabetes. Consequently, the quality of the rules produced using linguistic summarization is checked by the four quality measures namely: degree of truth, coverage, reliability and outliers. Among these, the degree of reliability is useful to find rules that represent dataset completely, and the outliers are used to find rules that deviate from the original result. Our experiment reveals results that are promising when compared to the PIMA dataset.**

Keywords—Fuzzy sets, type-2 (IT2), knowledge extraction, Linguistic summarization (LS), IF-THEN rules

I. INTRODUCTION

Linguistics is more than just a means of communication for humans. It influences our culture and thought processes. These are the semantics that give rise to a language which play a crucial role in the evolution of humans. This is because most of the information or data is communicated orally amongst humans, and they can convey an infinite number of ideas through language which can be built upon itself without limits. However, the communication between humans inherently consists of vagueness and uncertainties. This can be explained easily with the following example: if a person is asked "How is the weather outside", the person would usually reply using words such as "hot", "cold" etc. This introduces a good amount of uncertainty within the information that is being exchanged linguistically.

To handle such uncertainties, fuzzy sets (FS) [1] come to the rescue. The main advantage of employing FSs lie in their ability to capture linguistic uncertainties existing within linguistic expressions such as words. Zadeh introduced computing with words (CWW) [2] to efficiently handle linguistic uncertainties. CWW can thus be considered as an extension of classical mathematics, in a sense where linguistics is used instead of numerals for reasoning. It is typically used when the information mentioned involves linguistic features, such as words that are inherently vague.

As mentioned, information defined by humans is predominantly linguistic. Therefore, responses generated by humans are words, which in turn act as linguistic inputs to approaches based on CWW. Since machines do not understand words they are 'translated' to numeric data. These numeric data are then aggregated and retranslated to linguistic output in the form of recommendations. This is the basic scheme of CWW as proposed by Zadeh.

Some recent works within the domain of CWW include linguistic optimization solution methodologies based on CWW [3, 4], power optimization in handheld devices [5], decision making models [6, 7], and student strategy evaluation [37].

The demand for data is increasing exponentially in recent years. Therefore, the task of data summarization (DS) [8] is essentially required to provide brief and quick reviews of data without any manual analysis. According to the definition given by Mani and Maybury [9] "summarization is the process of distilling the most important information from the source (or sources) to produce an abridged version for a particular user (users) and task (or tasks)."

DS can be categorized into two main classes, namely: linguistic summarization (LS) and numerical summarization (NS) [10]. Mean, median and variance are the statistical aspects of data that are characterized under NS, but according to Yager summarization is useful if it gives summaries that are not terse as mean as well as treating the summaries of non-numeric data. This means that LS of data which gives summaries like "most of the students who study score good marks" or " if P is little and Q is small then L is tiny" is more conductive as it handles non-numeric data and is easily understandable. Linguistic information is much easier to remember rather than recalling numeric information. For example, "the boy is tall " is much easier to remember rather than " the boy is 150 cm". Linguistically expressed properties of objects are summarizers e.g. "tall girl" and are hence represented by FSs.

Different strategies of LS have been created over the previous years to adapt to the exponential amount of information generated and put away. There are two primary ways for programmed creation of LS: one, utilizing fuzzy logic tools [8], the other, with natural language generation (NLG) systems [11] [12]. Different techniques are introduced, based on fuzzy dependencies based on rules like "if the GDP is high, then the economy is growing". Systems based on these platforms are used to generate financial summaries [13]. Another technique used in fuzzy logic is conceptual trees [14] [15] in which the results are displayed as a hierarchy of concepts and not sentences. Another kind of LS is achieved through natural language generation (NLG) strategies, as already mentioned above. The main difference in both techniques is that in NLG there is more than one sentence and the data processing part is less detailed. However, NLG is less flexible as the selection of data is by an expert or is hardcoded.

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

In general, applying various algorithms on databases linguistic information could be obtained.

Multiple LS models [16] exist in the literature that utilize type-1 (T1) FSs for the representation of linguistic terms. However, T1 FSs are not capable of modeling the uncertainties associated with linguistic terms. Interval type-2 (IT2) FSs were employed as representational models for linguistic terms as they handle both the interpersonal and intrapersonal uncertainties associated with it [17]. Due to this, there has been a boost in research on T2 FSs [18]. Viewing the research trend for T2 FSs in [18]one can conclude that the field has matured enough and hence, finds applications to multiple domains such as multiple criteria group decision making [19], veracity handling and instance reduction in big data [20], multi-objective task scheduling in Industry 4.0 ecosystem [21] etc. Accordingly, the authors of [17] extended the idea of LS by utilizing IT2 FSs for IF-THEN rule generation. Taking motivation from the same, we utilize the model in [17].

Taking motivation from the same, we utilize the model in [17], to propose the idea of linguistic summarization (LS) based on IF-THEN rules and interval type-2 (IT2 FS) FSs in the context of diabetes. The dataset of concern is a collection of records of individuals corresponding to five crucial parameters in the domain of diabetes. This dataset is put together by posing questions to the concerned individuals. The application of diabetes is chosen because it is a health conditions that affects the world's most of the population. This makes the study of such a domain extremely important.

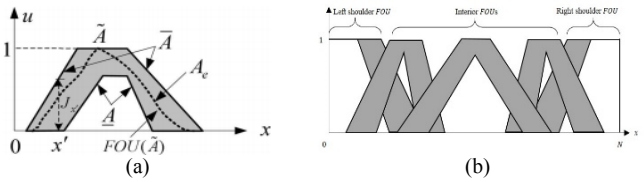
To facilitate the study mentioned above, numeric data of patients are collected on five parameters which are classified into their 3 respective classes. Further, the IT2 FS word models [22] are constructed for the dataset before the process of LS is carried out. Consequently, we calculate five quality measures i.e. truth value, coverage, reliability, outlier and simplicity to measure the different characteristics of IF-THEN rules produced from the collected dataset and present the top 10 rules of the dataset. These top 10 rules are the rules that have highest reliability i.e. have high coverage of the dataset and have high truth value. Thus these rules are the most useful ones in the dataset. Through this experiment, we observe that the quality of rules obtained through the process of LS is promising, when compared to the well-known PIMA dataset [23].

The rest of the paper is arranged as follows: Section II introduces preliminaries. Section III presents preliminary information about diabetes and the importance of the parameters selected to collect the dataset. Section IV introduces the detailed approach of linguistic summarization using IT2FSs, performed on the collected dataset of diabetes. Section V mentions the conclusion.

II. PRELIMINARIES

A. Linguistic summarization (LS)

A summary of a dataset consists of three parts viz. (1) a summarizer (S), (2) a quantity in agreement (Q) and, (3) a measure of validity (T). For e.g., consider that a dataset, $P = \{3,5,7,8,11,15\}$ is the set of marks obtained by students. We can discuss summaries in the form: $S =$ "$about\ 10$", $Q =$"$Most$" + "$...therefore\ most\ \ of\ the\ marks\ are\ about\ 10$". For different summarizers we obtain different values of T.



(a) (b)

Fig. 1 (a) Footprint of uncertainty (FOU), UMF and LMF. (b) Possible FOUs of words. [33]

The data should be summarized with the help of the summarizer (S). The summarizers can hold linguistic or numeric values. The linguistic values given to the summarizers depend on how those linguistic values are accommodated in FSs. For e.g., consider a set $F = \{1,3,4,6\}$. We can represent its summarizer as "near 6" and then accommodating it in a FS as $\{0/1, 0.5/3, 0.6/4, 1/6\}$.

The next part of LS is quantity in agreement (Q). This indicates the quantity to which the summarizer satisfies the data. The next part in LS is degree of truth that basically depicts the validity of the summary. It is calculated as follows:

$$T = Q(r), where\ r = \frac{1}{n} * \sum_{i=1}^{n} S(d_i)$$

$S(d_i)$ is the proportion of the dataset D which satisfies S.

B. Type -1 fuzzy sets (T1FS) [24]

Definition 1. A fuzzy set A is a crisp set of ordered pairs $\{x, \mu_A(x): x \in X\}$, where $\mu_A(x): X \rightarrow [0,1]$ is the membership value of A, and X is the universe of discourse. The support of A is expressed as:

$$Supp(A) =_{df} \{x \in X: \mu_A(x) > 0 \quad (1)$$

Definition 2. Cardinality of the T1FS is defined as follows

$$Card\ (A) = \sum_{x \in X} \mu_A(x) \quad (2)$$

C. Interval Type-2 FSs (IT2 FS)

Definition 3. An IT2 FS, $\tilde{A}$ is given by the MF $\mu_{\tilde{A}}(x,u)$, where $x \in$ X and $\mu_{\tilde{A}} \in J_x[0,1], i.e.$

$$\tilde{A}=\{((x,u),\mu_{\tilde{A}}(x,u) = 1) | \forall x \in X, \forall u \in j_x \subseteq [0,1]\} \quad (3)$$

where x is a primary variable with domain X; $\mu \in [0,1]$ which is the secondary variable, has domain $J_x \subseteq [0,1]$at each $x \in X$; J_x is primary membership of x and $\mu_{\tilde{A}}$ is the secondary grade of x which equals to 1 $\forall x \in X$and $\forall u \in J_x \subseteq [0,1]$. An e.g. of IT2FS can be viewed in Fig. 1(b).

Definition 4. Uncertainty about a fuzzy set is conveyed by the union of its all primary membership values and is called the footprint of uncertainty (FOU) (see Fig 1.(a)). It is represented as follows:

$$FOU\ (\tilde{A}) = \cup_{x \in X} j_x \quad (4)$$

where $\tilde{A}$ is a fuzzy set and j_x represents the primary membership value

Definition 5. The upper membership value (UMF) and the lower membership value (LMF) are T1 FS that form FOU as shown in Fig. 1(b).

$$J_x = [\mu_{\underline{A}}(x), \mu_{\overline{A}}(x)] \quad (5)$$

Therefore, Equation (4) can be expressed as follows

$$FOU\ (\tilde{A}) = \cup_{\forall x \in X} [\mu_{\underline{A}}(x), \mu_{\overline{A}}(x)] \quad (6)$$

TABLE I. QMS USED IN THIS PAPER IN CORRESPONDENCE TO HIROTA AND PEDRCYZ'S QMS.

Hirota and Pedrcyz's QMs	The QM used for the dataset
Validity	Degree of truth (T)
Generality	Degree of coverage (C)
Usefulness	Degree of reliability (R)
Novelty	Degree of outlier (O)
Simplicity	Degree of simplicity (S)

Definition 6. An embedded T1 FS is expressed as:

$$A = \int \frac{u}{x}, u \in J_x \quad (7)$$

where $\int$ means union.

D. Summaries with T1 FSs

According to Yager [3] the LS of data means to generate natural language sentences. A LS is of the form,

$$QC\ are/have\ S[T]$$

where Q is the quantity in agreement, C is the subject of the summary, S is the summarizer and T is the degree of truth. For example- "about 15% of the adults suffer from extreme stress", where "about 15%" is the quantity in agreement, "adults" is the subject and "extreme stress" is the summarizer. The antecedents and the consequents are determined once the dataset is mentioned. Antecedent and the consequent are specified by the user with their FS models After this, all rules are formed and quality measures (QM) are computed which holds various definitions.

To measure a quality of the summary, Hirota and Pedrycz [25] gave five features as mentioned below.

1) Validity: Summaries are extracted from data that has high confidence.

2) Generality: This feature depicts the support of the data to the summary.

3) Usefulness: This feature deals with the impact of the summaries on the user.

4) Novelty: This basically tells the outliers in the summary.

5) Simplicity: This feature depicts how simple a summary is i.e. the no. of antecedents and consequent the rule has.

The five QMs used in this paper corresponding to validity, generality, usefulness, novelty and simplicity are mentioned in Table I.

III. DIABETES AND PARAMETERS SELECTED

A. Natural History

The main cause of diabetes is insulin deficiency [26], and this can be due to pancreatic disorders, defects in the formation of insulin, destruction of cells, genetic defects, etc. The host factors [27] that are used in the prediction of diabetes are age, sex, obesity etc. Diabetes may happen at any age but studies [28] demonstrate that predominance rises steeply with age. In many countries the sex ratio having diabetes is about equal. Obesity characterized as an anomalous development of the fat tissue is expressed as BMI (body mass index). It is a major risk factor in diabetes because glucose is not used up by the cells which require energy to work. This leads to abnormal eating thus increasing the BMI. Genetic factors play an important role in the occurrence of diabetes [28]. If a person has a family history with diabetes he/she is at higher risk of having the disease in the future.

TABLE II. SAMPLE OF DATASET COLLECTED

Age (year)	BMI (kg/m^2)	FBS (mg/dL)	HB A1c (%)	Heredity	Condition
40	16	98	6.5	0.2	No
39	18	90	6.6	0.6	No
20	20	80	6.8	0.7	Pre
52	15	79	5.8	1.0	Pre
60	26	130	7.0	0.5	Pre
70	28	140	4.5	0.9	Dia
30	29	150	4	0.3	Pre
62	32	132	4.8	0.1	No
18	12	120	4.9	0.3	Pre
10	13	125	5.4	0.4	Dia

B. Screening for diabetes

Urine Examination: Urine test [29] for glucose after 2 hours of feast is utilized for distinguishing instances of diabetes. Most cases affirm that glucose is found in urine in extreme instances of diabetes but is found missing in milder forms of the disease. This is known as a lack of "sensitivity". Thus urine test is not considered appropriate for the case finding.

Blood Sugar Testing: Glucose levels in random blood samples give unsatisfactory results as they give a crude estimate of the frequency of diabetes. Fasting glucose value [30] alone is not considered to be reliable. Thus the 2-hour oral glucose may be used either alone or with the fasting value.

HBA1c (Hemoglobin A1ctest): For cases without diabetes, the normal range for the hemoglobin A1ctest level is between 4% and 5.6%. Hemoglobin A1c levels [31] between 5.7% and 6.4% mean you have a higher risk to be diabetic (Pre-diabetes). Levels of 6.5% or higher assures that you have diabetes.

Therefore, the five selected parameters on which the data is collected is "Age", "BMI", "Fasting blood sugar level" (FBS) and HBA1c test and Heredity. A sample of the collection of data (from 10 patients) is shown in Table II. The dataset was collected from 100 subjects who suffered from diabetes on the mentioned parameters through a survey. The dataset contains the values of the parameters along with the condition of the subject.

IV. LINGUISTIC SUMMARIZATION USING IT2FSS

Membership values of T1 FSs are formed by the preferences of experts. However, to model a linguistic term there should be two or more opinions, but using T1 FSs [1] is arbitrary as the membership value is crisp. Traditional fuzzy sets hold no strategies to handle such uncertainties.

LS using IT2 FSs [32] can handle both interpersonal and intrapersonal vagueness in the dataset [32]. The membership value of T1 FSs is crisp but the membership value of IT2 FS is specified in intervals which is useful in cases in which it is difficult to figure out an exact membership value for a FS.

A. Linguistic summarization using IF-THEN rules and IT2FSs

The IT2FS word models should be framed before linguistic summarization is carried out. This is done by the interval approach (IA) [33]**.** In this technique, a vocabulary of fuzzy terms is formed as shown in Table III and IV, and a group of people are asked the endpoints to the interval following the fuzzy word. After the preprocessing step (in which the outliers

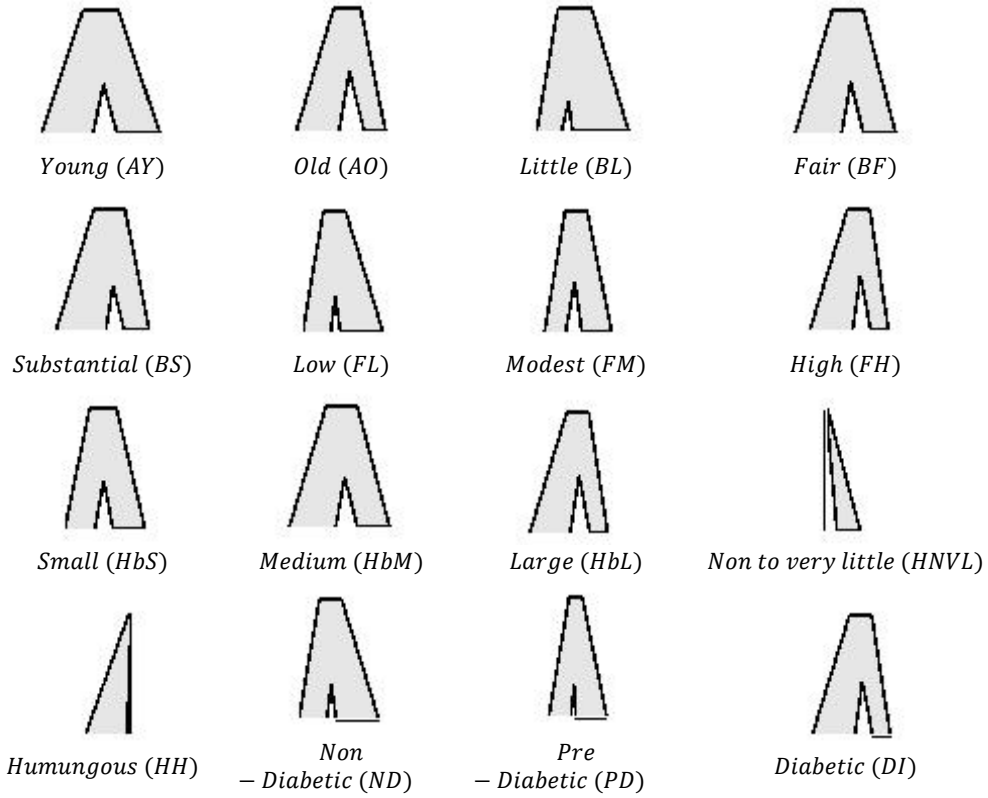


Fig. 2 Footprint of uncertainty (FOU) of the words obtained from interval approach on our dataset

TABLE III. PARAMETERS AND THERE FUZZY TERMS

Parameters	Range value	Fuzzy Term set
Age (year)	Less than 50	*Young* (*AY*)
	More than 50	*Old* (*AO*)
BMI (kg/m^2)	Less than 18.5	*Little* (*BL*)
	18.5-24.9	*Fair*(*BF*)
	More than 25	*Substantial* (*BS*)
FBS (mg/dL)	Less than 100	*Low* (*FL*)
	100-125	*Modest* (*FM*)
	More than 126	*High* (*FH*)
HbA1c results (%)	4% to 5.6%	*Small* (*HbS*)
	5.7% to 6.4%	*Medium* (*HbM*)
	More than 6.5%	*Large* (*HbL*)
Heredity	*Nonetolittle* (*HNVL*)	
	Humungous (*HH*)	

TABLE IV. CONDITION OF THE PATIENTS

Output Parameter	Possible recommendations
Condition of person	*No* diabetes: Non-diabetic (*No*)
	High risk of diabetes: Pre-diabetic (*Pre*)
	Assured presence of diabetes: Diabetic (*Dia*)

are eliminated) some intervals are eliminated and are mapped into the interior, left shoulder or right shoulder T1 FSs. Further the union of all T1 FSs is taken which results in a codebook of IT2 FSs as shown in Fig. 2.

The canonical form followed in generating IF-THEN rules [1] from the dataset is as follows.

$$IF\ a_1\ is\backslash has\ B_1, THEN\ a_2\ is\backslash has\ B_2\ \ [QM]$$

where B_1 and B_2 are words modeled by IT2FS and $QM \in [0,1]$, which indicates how useful a rule is.

B. Quality Measures for LS using IT2 FSs

The validity [34] represented as degree of truth, in LS is computed by the cardinalities of T1 FSs [2], to extend it further to IT2FS the following information is required.

Definition 7 [38]: The cardinality of the type-2 (IT2) FS is defined as:

$$C_D(\widetilde{B_1}) \equiv [C_D(\underline{B_1}), C_D(\underline{B_2})] = \sum_{m=1}^{M} \mu_{\underline{B_1}}(v_1^m), \sum_{m=1}^{M} \mu_{\overline{B_1}}(v_1^m)] \tag{8}$$

The average of the cardinality is defined as:

$$C_D(\widetilde{B_1}) = \frac{C_D(\underline{s_1}) + C_D(\overline{s_1})}{2} \tag{9}$$

Definition 8 [38]: The joint cardinality of type-2(IT2) $\{\widetilde{B_1}, \ldots..,\widetilde{B_N})$ is as follows:

$$C_D(\widetilde{B_1}\ldots.,\tilde{B}_N) \equiv [C_D(\underline{B_1}, \ldots\ldots, \underline{B_N}), C_D(\overline{B_1}, \ldots\ldots, \overline{B_N})] \tag{10}$$

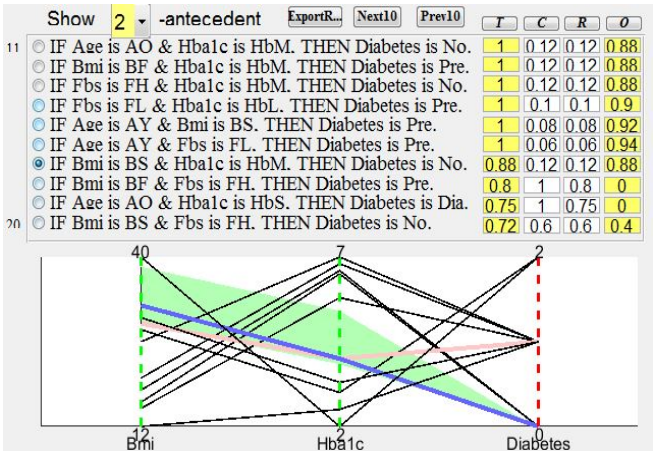
and the average of it is as follows:

$$C_D(\tilde{B}_1,\ldots,\tilde{B}_N) \equiv \frac{\left[C_D\left(\underline{B_1},\ldots\ldots,\underline{B_N}\right), C_D(\overline{B_1},\ldots\ldots,\overline{B_N})\right]}{2} \tag{11}$$

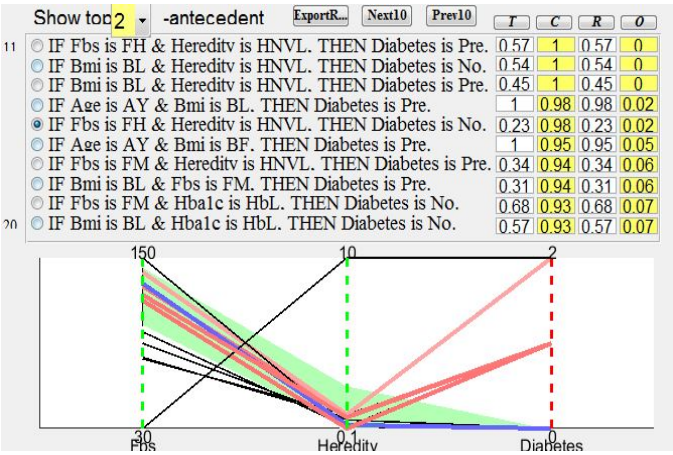
Using the cardinality and the joint cardinality the truth value can be represented as:

$$\tilde{T} = \frac{C_D(\widetilde{B_1},\widetilde{B_2})}{C_D(\widetilde{B_1})} \tag{12}$$

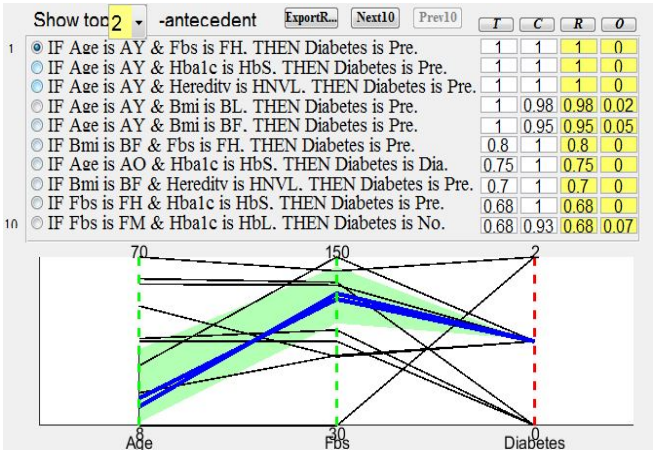
The generality of data [36] describes whether the rule supports enough data or not. Generality is not dependent on the value of truth. The formula for generality is as follows:



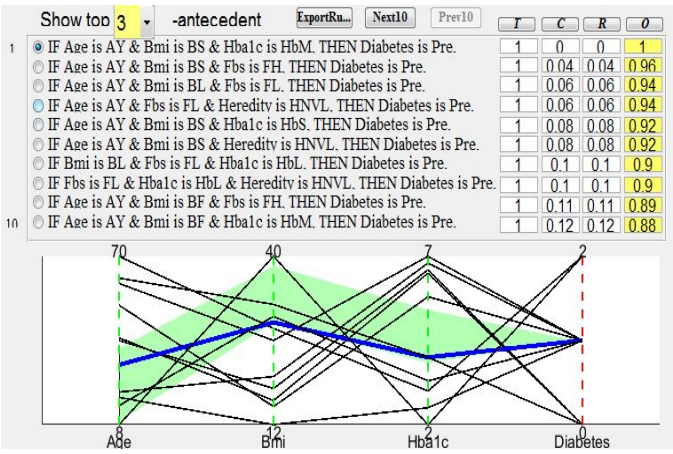
(a) Top 11-20 rules when T is the quality measure. It also depicts rule 17.



(b) Top 11-20 rules when C is the quality measure. It also depicts rule 15.



(c) Top 1-10 rules when R is the quality measure. It also depicts rule 1.



(d) Top 1-10 rules when O is the quality measure. It also depicts rule 1.

Fig. 3 Depiction of rules when different quality measures are considered.

The coverage ratio is calculated as:

$$r_c = \frac{\sum_{m=1}^{M} t_m}{M} \tag{13}$$

where t_m is calculated as follows:

$$t_m = \begin{cases} 1, \mu_{\overline{B_1}}(v_1^m) > 0 \; and \mu_{\overline{S_2}}(v_2^m) > 0 \\ 0, otherwise \end{cases} \tag{14}$$

After the coverage ratio is calculated the degree of sufficient coverage is calculated as follows:

$$C = f(r_c) where \; f function \; maps \; r_c \; into \; C \tag{15}$$

where, $f(r_c)$ is determined by 2 parameters r_1 and r_2

The degree of reliability [35] depicts how reliable a summary is, and can be calculated as:

$$R = min\,(T, C) \tag{16}$$

The degree of outlier [17] depicts how much a rule deviates from expected outcome.

$$O = \begin{cases} min(max(T, 1-T), 1-C), & T > 0 \\ 0 \qquad\qquad\qquad\quad , & T = 0 \end{cases} \tag{17}$$

The degree of simplicity is defined as follows:

$$S = 2^{2-l} \tag{18}$$

where l is the total number of antecedents and consequent used.

For multi-antecedent and multi-consequent rules, the degree of truth is calculated as follows:

$$T = \frac{C_D(\widetilde{B_1}, \ldots, \widetilde{B_n})}{C_D(\widetilde{B_1}, \ldots, \widetilde{B_k})} \tag{19}$$

and the coverage ratio r_c is calculated as:

$$t_m = \begin{cases} 1, \mu_{\overline{B_1}}(v_1^m) > 0 \; \forall_n = 1, \ldots, N \\ 0, otherwise \end{cases} \tag{20}$$

after r_c is obtained C is calculated as (15) and as C and T are crisp numbers R and O is are calculated as (16) and (17).

V. Linguistic Summarization of Diabetes Dataset

A MATLAB based GUI demonstrates IT2 FS based LS as given in [35]. We make use of this GUI to visualize the rules as well as their quality parameters on the collected dataset. The GUI finds the top 10 global rules of the dataset that has the maximum value of the degree of truth (T), degree of generality (C), degree of reliability (R) or degree of outlier (O). Clicking on the buttons of T, C, R, or O as shown in GUI updates the rules accordingly. The global rules are then displayed following the criteria chosen by the user. The user firstly has to select the number of antecedents and the program then displays the rules with all combinations of words of antecedents. The MATLAB based GUI can display up to 10 rules at a time [35], and accordingly, we show the top 10 rules here. A specific rule can be selected by clicking on the radio button adjacent to the displayed rule. To know the further rules

of the dataset the user can click on the NEXT10 button in the MATLAB based GUI that would display the next 10 rules of the dataset.

In GUI the blue lines represent those cases that follow the consequent and the degree of support depends on the intensity of the color. The red lines represent the outliers in the dataset and the power of violating the rule is equivalent to the intensity of the shade. The cases that are not necessary are shown in black color. The region that is covered by the rule is depicted in green color.

Fig. 3(a) depicts that if T is utilized as the ranking measure, a rule with high T may portray not very many cases; consequently, it is entirely conceivable that this standard portrays just outliers and, subsequently, cannot be believed. Like in rule 17th " If BMI is BS and HBA1c is HbM then Diabetes is NO" has T=0.88 but from Fig. 3(a) we can infer that only a few cases lie in the region described by the rule. Hence T alone cannot be relied upon for QM of LS.

In Fig. 3(b) when C is utilized as the positioning foundation, and if it has a high value it can depict a low level of truth, e.g. Like in rule 15th "If FBS is FH and Heredity is HNTVL then Diabetes is NO " has C = 1, implies numerous cases bolster this standard, yet from of Fig. 3(b), we notice, numerous cases disregard it as well (that is the reason its T = 0.23, which is a modest value). To be sure, this rule appears to be irrational. In this manner, C alone isn't a great QM either.

In Fig. 3(c) when R is utilized as the positioning model, a standard which has high R has a greater level of truth and adequate coverage, e.g. Like in rule 1st " If Age is AY and FBS is FH then Diabetes is Pre" has R=1, and from the figure in most cases antecedents follow the rule at contrasting degrees and henceforth, it portrays a valuable rule. Subsequently, R is a solid QM for LS.

In Fig. 3(d) if O is utilized as the positioning model, a rule which has high value of O depicts an exceptionally small number of cases e.g. consider the first rule with 3 antecedents" If Age is AY, BMI is BS, and HBA1c is HbM then Diabetes is Pre" has O = 1, from Fig. 3(d) we see that just one case follows this standard, which ought to be considered as exceptions. Consequently, O is valuable in finding irrelevant information and rules.

VI. Comparison with Rules Generated using PIMA Dataset

The PIMA dataset [23] consists of data from Indians suffering from Diabetes. This dataset consists of 768 instances with data for eight parameters namely: number of times the patient has been pregnant, glucose concentration in blood, blood pressure, triceps, serum insulin, BMI, pedigree, and age. The outputs to which the instances are classified are: non-diabetic and diabetic.

Authors of [17] utilize this dataset to perform LS to obtain IF-THEN rules. On the basis of the reliability measure (R), utilizing the PIMA dataset provides maximum reliable rules with a measure of 0.97 and an outlier measure of 0.02.

On the other hand, our constructed dataset consists of 250 instances of patient data for five attributes as described earlier. Our dataset classifies the instances into three classes: non-diabetic, pre-diabetic and diabetic. Utilizing this dataset for obtaining IF-THEN rules through LS using IT2 FSs results in the maximum reliable rule with a measure of 100 and an outlier measure of 0 (as shown in Fig. 3(c)). Therefore, our constructed dataset provides better rules with better accuracy considering lesser number of attributes. Also, three classes for mapping the condition of a patient is much more reliable.

VII. Conclusion

Summarization provides a means of answering questions about the collected datasets and it also formats the information for the analysts. Following the same lines, we have proposed a dataset with five attributes that are classified into 3 classes. Consequently, a MATLAB based GUI model was used to produce the top 10 rules that summarize the concerned dataset. The dataset used in the paper is better than the PIMA dataset of diabetes as in it there are 8 attributes with two classes [17] but the collected dataset has five attributes with three classes which gives more clarity to the rules of the disease as it works upon three classes. Reliability (R) and degree of outliers (O) are preferable QMs for LS over the degree of Truth (T) for the collected dataset. A high R recognizes a helpful rule with both adequate inclusion and degree of truth, while an excessive value of O recognizes outliers in data that requires further examination. Taking R as a QM for LS gives the top 10 rules (for diabetes) of the dataset collected, classified into 3 classes and taking 5 attributes into consideration.

References

[1] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338-353, 1965.

[2] M. Setnes, R. Babuska, and H. B. Verbruggen. "Rule-based modeling: Precision and transparency," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 1, pp. 165-169, 1998.

[3] P.K. Gupta, and P. K. Muhuri. “Perceptual reasoning based solution methodology for linguistic optimization problems,” *arXiv preprint* arXiv:2004.14933, pp. 1–14, 2020.

[4] P. K. Gupta, and P. K. Muhuri. “Extended Tsukamoto’s inference method for solving multi-objective linguistic optimization problems,” *Fuzzy Sets and Systems* vol. 377, pp. 102–124, 2019.

[5] P. K. Muhuri, P. K. Gupta, and J. M. Mendel, “Person Footprint of Uncertainty Based CWW Model for Power Optimization in Handheld Devices,” *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 3, pp. 558 - 568, 2020.

[6] T. Seth, and P. K. Muhuri, “A Linguistic Decision Making Model for Psychometric Tests,” *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, New Orleans, USA, pp. 1-6, 2019.

[7] T. Seth, and P. K. Muhuri, “Hesitant Fuzzy Linguistic Term Sets for Group Decision Making in Supplier Performance Evaluation,” *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Rio de Janeiro, Brazil, pp. 1-8, 2018.

[8] J. Kacprzyk, and R. R. Yager. “Linguistic summaries of data using fuzzy logic,” *International Journal of General System*, vol. 30, no. 2, pp. 133-154, 2001.

[9] K. Collins-Thompson, and J. Callan. "Automatic and human scoring of word definition responses." In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 476-483. 2007.

[10] J. Kacprzyk, A. Wilbik, and S. Zadrożny. “An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation,” *International Journal of Intelligent Systems*, vol. 25, no. 5, pp. 411-439, 2010.

[11] J. Kacprzyk, and S. Zadrozny. Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation,” *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 3, pp. 461-472, 2010.

[12] K. McKeown, J. Robin, and K. Kukich. “Generating concise natural language summaries,” *Information Processing & Management*, vol. 31, no. 5, pp. 703-733, 1995.

[13] B. Bouchon-Meunier, and G. Moyse. “Fuzzy linguistic summaries: Where are we, where can we go?, “ *In 2012 IEEE Conference on*

Computational Intelligence for Financial Engineering & Economics (CIFEr), pp. 1-8, IEEE, 2012.

[14] P. Mulhem, W. K. Leow, and Y. K. Lee, "Fuzzy conceptual graphs for matching images of natural scenes," In *Proceedings of the 17th international joint conference on Artificial intelligence,* vol. 2, pp. 1397-1402. 2001.

[15] A. Pigeau, G. Raschia, M. Gelgon, N. Mouaddib, and R. Saint-Paul. "A fuzzy linguistic summarization technique for TV recommender systems," *In The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03*, vol. 1, pp. 743-748, IEEE, 2003.

[16] J. M. Mendel, "Computing with words and its relationships with fuzzistics," *Information Sciences*, vol. 177, pp. 988–1006, 2007.

[17] D. Wu, and J. M. Mendel. "Linguistic summarization using IF–THEN rules and interval type-2 fuzzy sets". *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 1, pp. 136-151, 2010.

[18] A. K. Shukla, S. K. Banshal, T. Seth, A. Basu, R. John, and P. K. Muhuri, "A Bibliometric Overview of the Field of Type-2 Fuzzy Sets and Systems [Discussion Forum]," *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 89–98, Feb. 2020.

[19] T. Seth and P. K. Muhuri, "Type-2 Fuzzy Set based Hesitant Fuzzy Linguistic Term Sets for Linguistic Decision Making," *arXiv preprint* arXiv:2002.11714, pp. 1–19, 2020.

[20] A. K. Shukla, M. Yadav, S. Kumar, and P. K. Muhuri, "Veracity handling and instance reduction in big data using interval type-2 fuzzy sets," *Engineering Applications of Artificial Intelligence*, vol. 88, p. 103315, Feb. 2020.

[21] A. K. Shukla, R. Nath, P. K. Muhuri, and Q. M. D. Lohani, "Energy efficient multi-objective scheduling of tasks with interval type-2 fuzzy timing constraints in an Industry 4.0 ecosystem," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103257, Jan. 2020.

[22] F. Liu, and J. M. Mendel. "Encoding words into interval type-2 fuzzy sets using an interval approach". *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 6, pp. 1503-1521, 2008.

[23] J. A. Johnson, T. E. Nowatzki, and S. J. Coons. "Health-related quality of life of diabetic Pima Indians," *Medical care*, vol. 34, no. 2, pp. 97-102, 1996.

[24] G. Klir and B. Yuan. "Fuzzy sets and fuzzy logic," *New Jersey: Prentice hall*, 1995.

[25] K. Hirota, and W. Pedrycz. "Fuzzy computing for data mining," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1575-1600, 1999.

[26] R. De Marco, F. Locatelli, G. Zoppini, G. Verlato, E. Bonora, and M. Muggeo. "Cause-specific mortality in type 2 diabetes". *The Verona Diabetes Study*. *Diabetes care*, vol. 22, no. 5, pp. 756-761, 1999.

[27] D. Wark Boucher, K. Hayashi, J. Rosenthal, and A. L. Notkins. "Virus-induced diabetes mellitus. III. Influence of the sex and strain of the host." *Journal of Infectious Diseases*, vol. 131, no. 4, pp. 462-466, 1975.

[28] J. E Park. "Textbook of preventive and social medicine: a treatise on community health," *Banarsidas Bhanot*, 1972.

[29] R. G. Nelson, D. J. Pettitt, H. R. Baird, M. A. Charles, Q. Z. Liu, P. H. Bennett, and W. C. Knowler. "Pre-diabetic blood pressure predicts urinary albumin excretion after the onset of type 2 (non-insulin-dependent) diabetes mellitus in Pima Indians," *Diabetologia*, vol. 36, no. 10, pp. 998-1001, 1993.

[30] D. A. Sacks, W. Chen, G. Wolde-Tsadik, and T. A. Buchanan. "Fasting plasma glucose test at the first prenatal visit as a screen for gestational diabetes," *Obstetrics & Gynecology*, vol. 101, no. 6, pp. 1197-1203, 2003.

[31] L. S. Greci, M. Kailasam, S. Malkani, D. L. Katz, I. Hulinsky, R. Ahmadi, and H. Nawaz. "Utility of HbA1c levels for diabetes case finding in hospitalized patients with hyperglycemia," *Diabetes care*, vol. 26, no. 4, pp. 1064-1068, 2003.

[32] J. T. Starczewski, "Efficient triangular type-2 fuzzy logic systems," *International journal of approximate reasoning*, vol. 50, no. 5, pp. 799-811, 2009.

[33] J. M. Mendel, and D. Wu. "Perceptual computing: Aiding people in making subjective judgments," vol. 13. *John Wiley & Sons*, 2010.

[34] L. A. Zadeh. "Fuzzy logic, neural networks, and soft computing," *In Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh*, pp. 775-782, 1996.

[35] D. Wu, J. M. Mendel, and J. Joo. "Linguistic summarization using if-then rules," *In International Conference on Fuzzy Systems*, pp. 1-8. IEEE, 2010.

[36] J. B. Bowles, and C. E. Pelaez. "Application of fuzzy logic to reliability engineering," *Proceedings of the IEEE*, vol. 83, no. 3, pp. 435-449, 1995.

[37] P. K. Gupta, and P. K. Muhuri, "Computing with words for student strategy evaluation in an examination," *Granular Computing*, vol. 4, no. 2, pp. 167–184, Jun. 2018.

[38] I. K. Vlachos, and G. D. Sergiadis. "Subsethood, entropy, and cardinality for interval-valued fuzzy sets—an algebraic derivation," *Fuzzy Sets and Systems*, vol. 158, no. 12, pp. 1384-1396, 2007.