ELSEVIER

Fuzzy Sets and Systems 119 (2001) 329–335

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy equalization in the construction of fuzzy sets

Witold Pedrycz[a,b,*]

[a] *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Alta., Canada AB T6G 2G7*
[b] *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Received 14 December 1998; received in revised form 11 May 1999; accepted 8 July 1999

Abstract

In this note, we introduce and study a concept of fuzzy equalization. Fuzzy equalization concerns a process of building information granules that are semantically and experimentally meaningful. The experimental relevance of a given information granule (fuzzy set) is directly linked with an encapsulation of a certain experimental evidence conveyed by the respective probability density function of available data. We establish a detailed equalization algorithm developed for triangular fuzzy sets. The study elaborates on the role of the fuzzy equalization in system design. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Information granulation; Linguistic labels; Probability; Linguistic equalization; Triangular membership functions; Fuzzy contingency tables

1. Introductory remarks

There have been a lot of considerations dealing with the nature and usage of linguistic labels — fuzzy sets. The essence of all these discussions revolves around an origin of fuzzy sets. Where do they come from? The commonly encountered arguments point out at the origin of fuzzy sets viewed as the basic entities emerging in a process of human cognition and inherently associated with various problem solving activities. As a follow-up of some psychological findings, the number of fuzzy sets being used to granulate individual variables is often restricted to 7 ± 2 terms.

The current state of fuzzy modeling and simulation hinges on numeric data. First, these data are used to construct fuzzy models. Subsequently, numeric data are used to evaluate and verify the already developed structures. It is therefore highly justifiable to expect that all information granules used in system development need to be fully legitimized in terms of the experimental (numeric) data. This observation means that information granules used throughout the process should be both *semantically* meaningful and *experimentally* meaningful. The aspect of semantics of the individual fuzzy sets and their collections (families) has been already thoroughly discussed in the existing literature, see e.g. [2–4,7,8,10]. It exhibits a number of facets and concerns fuzzy sets to be complete and sufficiently disjoint. By subscribing to the data meaningfulness requirements, we become more cognizant about the important relationships between fuzzy sets and probability calculus. To rephrase this observation, one may ask about links between fuzzy sets and experimental data. Our conjuncture is that the linguistic terms emerge only if there is an experimental

[*] Tel.: +1-204-474-8380; fax: +1-204-261-4639.
E-mail address: pedrycz@ee.ualberta.ca (W. Pedrycz).

PII: S0165-0114(99)00135-9

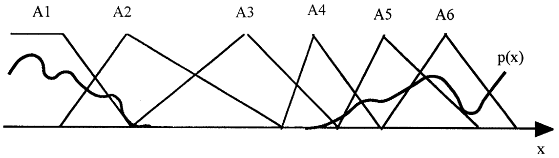


Fig. 1. Experimental data vis-à-vis linguistic labels; note that there is no justification as to the existence and specific distribution of some fuzzy sets (such as A_3 and A_4).

evidence behind them. Naturally, if there are no pertinent experimental data, there is no point of constructing a linguistic label: its existence cannot be justified in terms of numeric data. Moreover, one cannot expect this linguistic label to be verifiable in the framework of any fuzzy construct based on some previously existing experimental evidence. If introduced, such label accomplishes nothing. It does not encapsulate any data. Fig. 1 alludes directly to this point by showing several linguistic labels superimposed over a collection of the available experimental data. Observe that some of the linguistic labels do not deal with any experimental evidence. The form and distribution does not bear any relevance.

Let us also make another observation that becomes quite evident. With the advent of neural networks and various ideas of neurofuzzy computing [1,5], it is important to cast the question of the origin of the linguistic labels in this specific conceptual and algorithmic framework. It is usually stated that neural networks (or being more precise fuzzy neural networks) can construct fuzzy sets on their own. Despite the particular learning techniques being used to train the network (including the parameters of the underlying fuzzy sets), it is also evident that behind any changes or build-up of the fuzzy sets stand numeric data. The data show up quite explicitly in the form of the performance index we minimize. In this way, the data set affects the form of the fuzzy sets. The changes of the membership functions are also made legitimate once supported by the existing data. This reinforces the previous argument being made in the context of data illustrated in Fig. 1.

The underlying objective of this study is to develop fuzzy sets defined in a certain universe of discourse in such a way that they are data-*justifiable.* Put the problem in a different perspective: when looking at the experimental numeric data we are involved in their granular (fuzzy) *equalization.* This equalization is completed in the form of fuzzy sets. By pursuing this avenue, it becomes apparent that probability and fuzziness attempt to collaborate rather than compete [9].

With the reasons outlined later on, in this paper we confine ourselves to triangular fuzzy sets with $\frac{1}{2}$ overlap occurring between two successive linguistic terms. Through the $\frac{1}{2}$ overlap realized there, all these fuzzy sets form a fuzzy partition of the universe of discourse $\mathbf{X}$. Furthermore (which is dominant in existing applications), we assume that $\mathbf{X}$ is just a subset of reals ($\mathbf{X} \subset \mathbf{R}$).

The material is organized as follows. First, we formulate the problem in Section 2. The complete algorithm is included in Section 3. This is followed by illustrative experimental examples covered in Section 4.

2. Linguistic data equalization

The concept of granular data equalization originates from the idea of fuzzy events [6]. Any fuzzy set defined in some universe of discourse over which given is also some probability density function (pdf) either in its continuous or discrete format, comes with some cumulative probability. This probability is determined by integrating over the support of the fuzzy set. More precisely, we obtain

$$P(A) = \int_x A(x)p(x)\,\mathrm{d}x, \tag{1}$$

where A is the fuzzy set of interest whereas $p(x)$ denotes the corresponding pdf defined in $\mathbf{X}$.

Usually when discussing fuzzy sets over $\mathbf{X}$ we are concerned with a family of fuzzy sets, say $\mathbf{A} = \{A_1, A_2, \ldots, A_c\}$. To make each A_i meaningful, we require that

$$P(A_1) = P(A_2) = \cdots = P(A_c) = 1/c. \tag{2}$$

The above equalization condition (1) states that each element of $\mathbf{A}$ carries (encapsulates) the same amount of experimental evidence (see also Fig. 2). It is noticeable that fuzzy sets become more specific in the regions of $\mathbf{X}$ where pdf attains local minima. On the other hand, in the areas of low values of pdf we

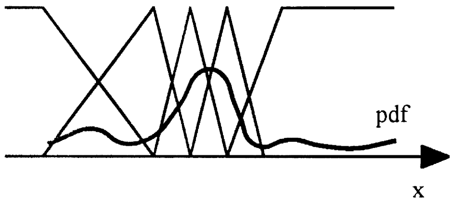


Fig. 2. The idea of linguistic equalization.

need fuzzy sets of broader support to gain sufficient evidence.

For the continuous pdf, the calculations of $P(A_i)$ can be carried out following (1). For the discrete data set $\mathbf{X}=\{x_1,x_2,\ldots,x_N\}$, this formula is reformulated as follows:

$$P(A)=\frac{1}{N}\sum_k A(x_k).$$

Several obvious, yet interesting observations can be made, refer again to Fig. 2:

- for the uniform pdf, the required fuzzy sets of **A** are all equal in terms of their energy measure of fuzziness. This measure is given in the form

$$\int_x A(x)\,\mathrm{d}x=\frac{1}{c}, \tag{3}$$

- fuzzy sets defined over the regions of **X** with higher probability values become more specific (their measure of fuzziness gets lower),
- with the increased number of the fuzzy sets in **A** their probabilistic evidence diminishes.

The formulation of the problem, as conveyed by (2), is concise and straightforward. The solution to it may be quite complicated depending on the form of the membership functions of the underlying pdfs. Our intent is to treat the linguistic equalization as a basic vehicle for data preprocessing completed in the realm of fuzzy sets. Thus, it is justifiable to confine to a certain family of fuzzy sets **A** so that this helps reduce the ensuing computational effort.

3. The algorithm

The algorithm outlined below realizes the idea of data equalization through a series of triangular fuzzy numbers with an $\frac{1}{2}$ overlap between successive fuzzy sets. Moreover, the first as well as the last fuzzy set in **A** is defined by a trapezoidal membership function. These assumptions as to the family of the fuzzy sets used here (even though they may look quite restrictive) are often encountered in practice. They are also well justified taking into consideration an algorithmic substance of the proposed method. The way of equalization is the following:

0. Specify the number of elements of linguistic granules in **A**, say "c".
1. Start from the lower bound of **X** denoted by $x_{\min}$.
2. Proceed towards higher values of **X** computing the moving value of the integral

$$\int_{x_{\min}}^{a} A_1(x)p(x)\,\mathrm{d}x=\int_{x_{\min}}^{a} p(x)\,\mathrm{d}x$$

   (this integral describes the characteristic part of the membership function of A_1). Stop once the value of this integral has reached the value of $\frac{1}{2}c$ and record the corresponding value of the argument. Denote the value of this argument by "a".
3. We determine the upper bound (b) of the support of A_1 so that the probability of the fuzzy event implied by the decreasing part of the membership function becomes equal to

$$\int_a^b A_1(x)p(x)\,\mathrm{d}x=\frac{1}{2c}. \tag{4}$$

4. For the triangular fuzzy sets (starting from A_2 and proceeding with A_3, A_4, etc.) we compute the probability of the fuzzy event associated with the increasing part of the fuzzy set

$$\varepsilon=\int_a^b A_2(x)p(x)\,\mathrm{d}x. \tag{5}$$

5. Next, we optimize the decreasing part of the membership function by determining the upper bound of support of the fuzzy set such that it satisfies the condition

$$\int_b^c A_2(x)p(x)\,\mathrm{d}x=\frac{1}{c}-\varepsilon \tag{6}$$

   (for details refer to Fig. 3).
6. Repeat steps 4 and 5 for the successive fuzzy sets in **A**.

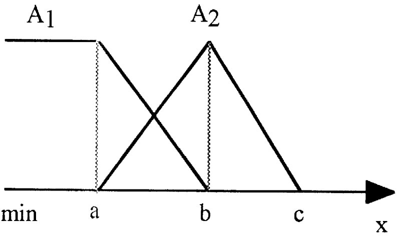


Fig. 3. The concept of linguistic equalization — a summary of pertinent optimization details.

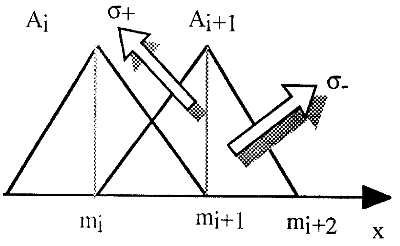


Fig. 4. Realizability condition of fuzzy equalization.

There are several reasons behind the use of the triangular fuzzy numbers:

- Triangular fuzzy sets with the $\frac{1}{2}$ overlap produce an error-free decoding. This means that once any numeric datum has been transformed (encoded) via these fuzzy sets giving rise to the corresponding membership values, such membership grades along with the corresponding modal values of the linguistic terms help decode the original datum without any error. This feature is commonly utilized when interfacing fuzzy set constructs with a numeric world. It is worth underlining that any departure from the $\frac{1}{2}$ overlap leads to the nonzero decoding error.
- These fuzzy sets form a fuzzy partition of **X** meaning that for each element in **X** the sum of the membership values is equal to 1. This helps define the level of probabilistic support to be associated with the individual fuzzy set.

The algorithm is straightforward. There is, however, a certain requirement one has to be aware of to make the construct (fuzzy sets) meaningful. It arises as an effect of a two-phased development of the individual fuzzy sets of **A** (refer to Fig. 4).

As fuzzy set A_i has been already determined, this predetermines the probability of the positive (increasing) segment of A_{i+1}. Denote it by σ_+,

$$\sigma_+ = \int_{m_i}^{m_{i+1}} A_{i+1}(x)p(x)\,\mathrm{d}x,$$

what is left to determine the negative (decreasing) segment of A_{i+1}. It comes with the probability σ_- computed as

$$\sigma_- = \frac{1}{c} - \sigma_+.$$

Obviously, if σ_+ exceeds $1/c$ then A_{i+1} cannot be constructed as no positive probability cannot be assigned to the negative segment of the fuzzy set.

The realizability requirement assuring the coherency of the overall design can be formulated as follows:

$$\int_{m_i}^{m_{i+1}} A_{i+1}(x)p(x)\,\mathrm{d}x < \frac{1}{c}$$

that holds for any linguistic term of **A**.

One should stress that the algorithm in its current version is geared toward the granulation of a single variable. It cannot be extended to a multivariable case in a straightforward manner. The main reason behind this lies in the linear ordering of the modal values of the fuzzy sets where this order is required to carry out all computing. Nevertheless, the generalization to the multivariable case could be easily completed by taking a bottom-up development approach. We simply consider each variable separately, develop fuzzy sets therein, and finally aggregate these fuzzy sets in the form of the respective fuzzy relations. The combination operator (Cartesian product) merging the individual fuzzy sets can be realized by taking any t-norm. This way of pursuing the development of the fuzzy constructs emphasizes the design of the basic entities, namely fuzzy sets rather than focuses on their composites (fuzzy relations). To make the picture complete, one should emphasize that the effect of an interaction between fuzzy sets in any multidimensional problem has not been fully addressed in an explicit manner. The only vehicle available at this point comes with the choice of the specific t-norm that can help model a strength of interaction between the contributing fuzzy sets. The aspect of an experimental verification of this facet is left open.

4. Experimental examples

The experiments reported in this section deal both with a continuous pdf as well as experimental data (resulting in discrete histograms).

The continuous case involves a Gaussian pdf that is eventually the most commonly used form of the probability distribution function. This pdf reads as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

In the experiment, we assume that $m = 3.0$ and $\sigma = 1.0$. The universe of discourse $\mathbf{X}$ is taken as a segment of real numbers spreading from 1 to 6, $\mathbf{X} = [1, 6]$. The parameters of the optimized triangular membership functions are given below

1.0000
1.8375
2.5971
3.0097
3.6265
4.2585

In the sequel, the resulting membership functions are portrayed in Fig. 5. Observe that the boundary fuzzy sets are far broader (less specific) that the rest of the linguistic terms. This is quite intuitive and is caused by the long yet quite limited "tails" of the Gaussian distribution. To compensate for low mass of probability in these two regions, we need fuzzy sets of lower granularity (broader support).

In the second experiment, the discrete probabilities are provided by experimental data that form a part of the housing data set available at the UC at Irvine (UCI Repository of Machine Learning Databases and Domain Theorem, http://www.ics.uci.edu). We take the last variable of the data set that concerns price of real estate (given in K$). The produced histogram involves 506 data points and is visualized in Fig. 6.

Note that the histogram is quite distinct from the standard Gaussian distribution and exhibits a number of local modal values.

To complete a comprehensive analysis, we carry out fuzzy equalization for several number of the local linguistic terms. The results are shown in a tabular form (Table 1).

In the case of $c = 4$ and 6, the resulting fuzzy sets are given in Fig. 7.

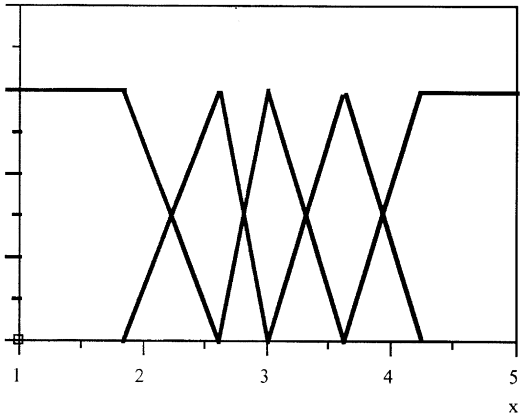


Fig. 5. Membership functions of the triangular fuzzy sets resulting from the Gaussian probability function.

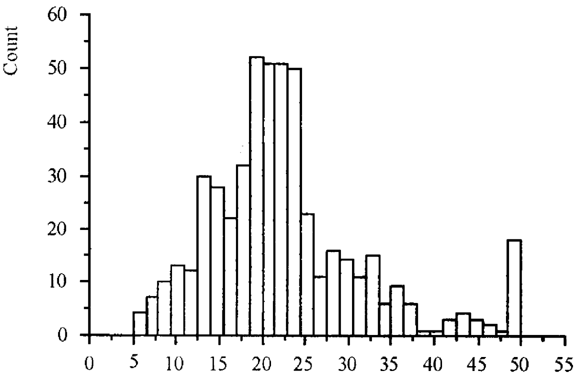


Fig. 6. A histogram of real estate price (in thousand of $).

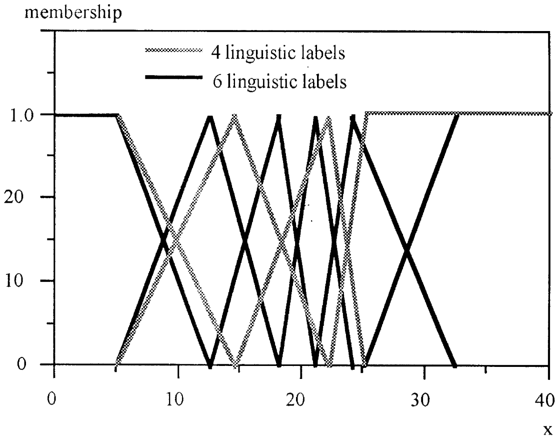


Fig. 7. Membership functions of the triangular fuzzy sets completing fuzzy equalization for $c = 4$ and 6.

Table 1
Modal values of the triangular fuzzy sets for different values of "c"

$c=3$	$c=4$	$c=5$	$c=6$	$c=7$	$c=8$
5.00	5.00	5.00	5.00	5.00	5.00
14.45	13.44	12.71	11.92	11.37	10.85
22.18	20.02	18.21	17.15	16.19	15.60
25.24	22.45	21.27	20.23	19.31	18.13
	30.44	24.10	22.17	20.92	20.60
		32.42	25.27	23.61	21.74
			35.28	25.30	24.32
				38.70	26.41
					40.18

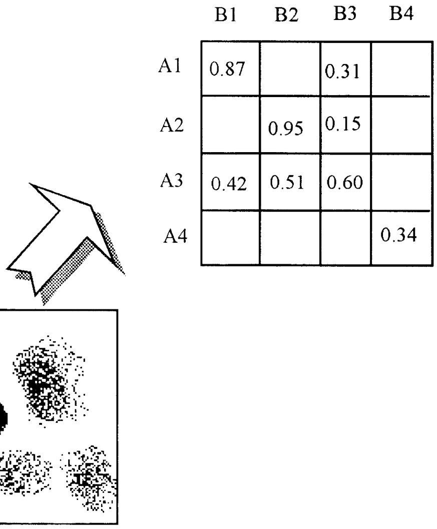
5. Fuzzy equalization in system design

In a nutshell, fuzzy modeling and system design arising within this setting emerge as a synonym of perceiving and capturing dependencies between information granules (regarded as fuzzy sets or fuzzy relations). Rather than being interested in minute and quite often irrelevant details, the focal point is to reveal dependencies (associations) at the level of some meaningful and easily comprehensible chunks of information. Two issues should be stressed as to the current practices of modeling and design exploiting information granules:

- very often fuzzy models are built up to very detailed relationships (dependencies) between information granules;
- fuzzy models are constructed (and verified afterwards) based on experimental (and commonly numeric) data.

Surprisingly enough, not too much attention has been paid to the articulation of the relevance of the information granules using which all constructs are developed. The underlying requirement of viewing all such granules legitimate (that is supported by experimental data to the same extent) sounds like a solid modeling prerequisite. This is nothing but an alternative formulation of the fuzzy equalization. Subsequently, the system design becomes apparently a two-phase procedure consisting of the following phases:

- ○ the construction of meaningful linguistic granules — conceptual tidbits;
- ○ the development of the detailed fuzzy model being dwelled on the already specified linguistic granules.



	B1	B2	B3	B4
A1	0.87		0.31	
A2		0.95	0.15	
A3	0.42	0.51	0.60	
A4				0.34

Fig. 8. An example of the fuzzy contingency table constructed for two variables with a number of linguistic granules defined for the individual variables and resulting in the families of fuzzy sets **A** and **B**.

The first phase predetermines an overall performance of the detailed model. With the equalized linguistic terms, the resulting model gets more balanced. Obviously, the impact of the equalization itself could vary from case to case and depend on the complexity of the relationships to be described by the particular model. The less detailed and sophisticated the ensuing relationships are, the more crucial the phenomenon of the linguistic equalization. In a simple scenario of a fuzzy contingency table (see Fig. 8) this impact becomes profoundly visible.

As the linguistic granules are balanced, the entire table is also equalized meaning that each entry of the table contributes to the model to a similar degree. Put it differently: as there is the same level of experimental (numerical) evidence behind each entry of the table, the strengths of the associations between any two entities can easily be compared. Note also that the effect of a highly unbalanced model could produce somewhat misleading conclusions. For instance, the level of association could be high between two information granules yet these two information granules may not be strongly supported by the experimental data. As a consequence, the otherwise high degree of association is not overly meaningful.

6. Conclusions

We have studied the concept of fuzzy equalization regarded as a basic vehicle of construction linguistic labels that are both semantically and experimentally meaningful. The way of building fuzzy sets underlines an important synergy between the technology of fuzzy sets and probability theory. The detailed algorithm is provided for triangular fuzzy sets. We have also identified the pertinent realizability conditions. The selection of this class of membership functions is strongly supported by the current system design practices. Finally, the study elaborates on the impact the equalization effect has on system design.

Acknowledgements

Support from the Natural Sciences and Engineering Research Council of Canada (NSERC) is gratefully acknowledged.

References

[1] IEEE Trans. Neural Networks — Special Issue on Fuzzy Logic and Neural Networks 3 (1992).

[2] G.J. Klir, T.A. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] G.J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[4] W. Pedrycz, F. Gomide, An Introduction to Fuzzy Sets: Analysis and Design, MIT Press, Cambridge, MA, 1998.

[5] H. Takagi, I. Hayashi, *NN*-driven fuzzy reasoning, Internat. J. Approx. Reason. 5 (1991) 191–212.

[6] L.A. Zadeh, Probability measures of fuzzy events, J. Math. Anal. Appl. 22 (1968) 421–427.

[7] L.A. Zadeh, Fuzzy sets and information granularity, in: M.M. Gupta, R.K. Ragade, R.R. Yager (Eds.), Advances in Fuzzy Set Theory and Applications, North-Holland, Amsterdam, 1979, pp. 3–18.

[8] L.A. Zadeh, The role of fuzzy logic in the management of uncertainty in expert systems, Fuzzy Sets and Systems 11 (1983) 199–227.

[9] L.A. Zadeh, Probability theory and fuzzy logic are complementary rather than competitive, Technometrics 37 (1995) 271–276.

[10] H.J. Zimmermann, Fuzzy Set Theory and Its Applications, 2nd Edition, Kluwer, Boston, MA, 1991.