Advances in Artificial Intelligence and Machine Learning; Research 1 (1) 86-93 Received 10-06-21; Accepted 25-06-21; Published 02-07-2021

Why Cauchy Membership Functions: Efficiency

Javier Viaña, Stephan Ralescu, {VIANAJR,RALESCS}@UCMAIL.UC.EDU
Kelly Cohen, and Anca Ralescu {COHENKY,RALESCAL}@UCMAIL.UC.EDU
University of Cincinnati,
Cincinnati, OH 45219, USA.
Vladik Kreinovich VLADIK@UTEP.EDU
University of Texas at El Paso, El Paso, TX 79968, USA.

Corresponding Author: KellyCohen, Vladik Kreinovich, Anca Ralescu.

Abstract

Fuzzy techniques depend heavily on eliciting meaningful membership functions for the fuzzy sets used. Often such functions are obtained from data. Just as often they are obtained from experts knowledgeable of the domain and the problem being addressed. However, there are cases when neither is possible, for example because of insufficient data, or unavailable experts. What functions should we choose and what should guide such choice? This paper argues in favor of using Cauchy membership functions, thus named because their expression is similar to that of the Cauchy distributions. The paper provides a theoretical explanation for this choice.

Keywords: Fuzzy sets, membership functions, Cauchy membership function.

1. INTRODUCTION

It is well known that the introduction of fuzzy sets opened up new possibilities in modeling and reasoning under uncertainty and imprecision [1]. The introduction of the notion linguistic variable has brought about additional benefits allowing for quantitative computation and its interpretation in word and reasoning [2].

In many practical applications of fuzzy techniques (see, e.g., [3-7,1]), the membership functions can be obtained from the experts. In other applications, the fuzzy sets are elicited directly from the data without the intervention of a human expert, imposing some condition on the underlying summarization procedure [8,9]. However, an important question arises: what is to be done when neither of these two approaches can be used? What functions should we then use? Experiments (see, e.g., [10,11]) show that in many applications, the following membership functions work best:

$$\mu_x(x) = \frac{1}{1 + \frac{(x-a)^2}{k^2}}. \tag{1}$$

Citation: Javier Viaña, et al. Why Cauchy Membership Functions: Efficiency. Advances in Artificial Intelligence and Machine Learning. 2021;1(1):6.

The expression (1) describing these membership functions is similar to the known expression for the probability density function $f(x)$ of a Cauchy distribution (see, e.g., [12]):

$$f(x) = const \cdot \frac{1}{1 + \frac{(x-a)^2}{k^2}}. \tag{2}$$

Because of this similarity, membership functions (1) are known as *Cauchy* membership functions.

A natural question is: how can we explain this empirical fact – that Cauchy membership functions work better than other functions tried? To answer such a question, we must define in a precise manner what it is really meant by "work better" and it is here suggested that this may be defined from two points of view – efficiency and reliability. In the remainder of the paper the efficiency is considered, showing that along with Gaussian membership functions, the Cauchy membership functions lead to efficient learning. To clarify, by efficient we mean an approach where, in the process of training a fuzzy model, straightforward one step computations are possible. This paper explores theoretically (in an almost axiomatic-like manner, that is, by imposing desired properties on the membership function) the basis for selection of a membership function in order to achieve efficient learning.

2. WHICH MEMBERSHIP FUNCTIONS LEAD TO THE MOST EFFICIENT LEARNING

2.1 Formulation of The Problem

From expert rules to fuzzy rules. One of the main reasons why Lotfi Zadeh invented fuzzy techniques was to translate expert rules that use imprecise (fuzzy) natural-language properties like *small, medium*, etc., into a precise control strategy. For this purpose, to each such property P, Zadeh proposed to assign a function $\mu_P(x)$ (known as membership function) that describes, for each possible value x of the corresponding quantity, the degree to which, according to the expert, an object with this value satisfies the property P – e.g., to what extent the amount x is small. This degree is usually assumed to be from the interval [0,1].

This is how first applications of fuzzy techniques emerged: researchers elicited rules and membership functions from the experts, and used fuzzy methodology to design a control strategy. The resulting control was often reasonably good, but not perfect. So, a natural idea was proposed: to use the original fuzzy control as a first approximation, and then to tune its parameters based on the practical behavior of the resulting system.

This *fuzzy learning* idea was first used in situations when we have expert rules that provide a reasonable first approximation. However, it turned out that this learning algorithm leads to a reasonable control even in the absence of expert rules, i.e., based solely on data.

Natural question: which membership function should we use? When starting with expert knowledge, membership functions are elicited from the experts. But when using fuzzy learning to situations when there is no expert knowledge, a natural question is: which membership functions should we use?

2.2 How to Select Membership Functions

Main idea: A natural idea is to select a membership function that would make learning faster. How can we do that?

Need for differentiability: The main objective of any learning is to optimize the corresponding objective function – a loss or cost function, which measures the discrepancy between the desired and actual behavior of the system. That is, if we have examples of desired outputs, then the objective is to minimize the discrepancy between the values produced by the system and the values we want to obtain.

Since the invention of calculus, the most efficient optimization techniques are based on computing the derivatives: one of the main objectives (and still one of main uses) of calculus is to identify points where a function attains its maximum or minimum among the roots of the first derivative. In machine learning, one of the simplest approaches is to achieve this is based on *gradient descent*.

The result of processing by several fuzzy layers is a composition of functions corresponding to each layer. So, to compute the derivative of the resulting transformation, we need to know the derivatives corresponding to each layer. From this viewpoint, to find a membership function that will make learning faster, we need to find membership functions which are differentiable and whose derivatives are easy to compute. Ideally, it should be possible to express such derivatives in terms of the original function (as is the case, for example, for the sigmoid and hyperbolic tangent functions often used in training neural networks).

In more precise terms, the problem is as follows: when computing the derivative $\mu'(x)$ for some input x, it is desired to use the fact that $\mu(x)$ has already been computed. Thus, in computing the value $\mu'(x)$,we can use not only the input x, but also the value $\mu(x)$. In other words, we are looking for an expression $\mu'(x) = f(\mu(x), x)$ for the simplest possible function of two variables, $f(a, x)$.

The meanning of simplest: In a computer, the only hardware supported operations with numbers are arithmetic operations: addition, subtraction (which, for the computer, is, in effect, the same as addition), multiplication, and taking an inverse (division is implemented as $\boldsymbol{a}/\boldsymbol{b} = \boldsymbol{a} \cdot (\boldsymbol{1}/\boldsymbol{b})$). To be more precise, computing an inverse is also implemented as a sequence of additions, subtractions, and multiplications, so each computation actually consists of additions, subtractions, and multiplications – and thus, computes a polynomial, since a polynomial can be defined as any function that can be obtained from variables and constants by using addition, subtraction, and multiplication. For example, to compute $exp(x)$ or $sin(x)$, most compilers compute the value of a polynomial that approximates the desired function – usually this polynomial is simply the sum of the first few terms of this function's Taylor expansion.

From this viewpoint, looking for the simplest function $f(a, x)$ means looking for a polynomial $f(a, x)$ that can be obtained by using the smallest possible number of arithmetic operations. (In a computer, unary minus is easy, so it is not counted.) Moreover, it is customary in machine learning (e.g., in regression problems) to look for the smallest degree polynomial to order to avoid overfitting.

Required asymptotic behavior: A typical membership function corresponding to notions like small and medium is only satisfied, with a reasonable degree, for a bounded set of values. That is, the support of a membership function is bounded. Thus, in the limits, when $\boldsymbol{x} \to \infty$ or $\boldsymbol{x} \to -\infty$, we

should have $\boldsymbol{\mu(x) \to 0}$. Thus, it makes sense to consider membership functions with this asymptotic property.

More over, most membership functions do not just asymptotically tend to 0, they are equal to 0 outside some intervals. For such function, in the areas where $\mu(x) = 0$, we expect $\mu'(x) = 0$, i.e., we have $f(0, x) = 0$ for all x. Since the function $f(a, x)$ is a polynomial, this means that all its monomials must be proportional to a, i.e., we must have $f(a, x) = a \cdot g(a, x)$ for some function $g(a, x)$. Thus, looking for the simplest function $f(a, x)$ means looking for the simplest functions $g(a, x)$. The cases when computing $g(a, x)$ requires 0 or 1 arithmetic operation are considered next.

When computing $\boldsymbol{g(a, x)}$ requires no arithmetic operations: This means that the value $\boldsymbol{g(a, x)}$ is equal to one of the given values, i.e., to $\boldsymbol{g(a, x) \in \{a, x, c\}}$ for some constant $\boldsymbol{c}$.

- If $g(a, x) = a$, it follows that $\mu'(x) = f(\mu(x), x) = \mu(x) \cdot g(\mu(x), x) = \mu(x) \cdot \mu(x) = \mu(x)^2$, i.e., $\frac{d\mu(x)}{dx} = \mu(x)^2$ hence $\frac{d\mu(x)}{\mu(x)^2} = dx$. Integrating, we obtain $-\frac{1}{\mu(x)} = x + C$, and hence $\mu'(x) = -\frac{1}{x+C}$. This function is unbounded, so it cannot serve as a membership function. In this case, adding unary minus, i.e., considering $g(a, x) = -a$, does not help.
- If $g(a, x) = x$, it follows that $\mu'(x) = \mu(x) \cdot x$, i.e., $\frac{d\mu(x)}{dx} = \mu(x) \cdot x$ hence $\frac{d\mu(x)}{\mu(x)} = x \cdot dx$. Integrating, obtains $\ln(\mu'(x)) = \frac{x^2}{2} + C$, i.e., $\mu'(x) = A\exp\left(\frac{x^2}{2}\right)$ for some constant $A = \exp(C)$. This is not a membership function, but by adding unary negation, i.e., by considering $g(a, x) = -x$, we obtain $\mu'(x) = \exp\left(-\frac{x^2}{2}\right)$ – a very reasonable case of Gaussian membership functions.
- If $g(a, x) = c$, it follows that $\mu'(x) = c \cdot \mu(x)$, i.e., $\frac{d\mu(x)}{dx} = c \cdot \mu(x)$ hence $\frac{d\mu(x)}{\mu(x)} = c \cdot dx$. Integrating, we obtain $\ln(\mu'(x)) = c \cdot x + C$, i.e., $\mu'(x) = A \cdot \exp(c \cdot x)$ – also not membership functions.

When computing $\boldsymbol{g(a, x)}$ requires one arithmetic operation: This operation can be addition/subtraction or multiplication.

1. For addition, we can have $g(a, x) = a + a$, $g(a, x) = a + c$, $g(a, x) = x + x$, $g(a, x) = x + c$, or $g(a, x) = a + x$. In the first case, leads to an unbounded function. The second case, leads to a sigmoid function – that does not have the right asymptotic behavior for $x \to \pm\infty$. The third and fourth cases, lead to the Gaussian functions – re-scaled in the third case and shifted in the fourth case. Finally, the last case leads to a reasonable differential equation $\mu'(x) = \mu(x) \cdot (\mu(x) + x)$, but the problem is that this equation does not have an explicit solution. This means that while computing $\mu'(x)$ using $\mu(x)$, computing $\mu(x)$ itself will be difficult – so this case should also be dismissed.
2. For multiplication, the same five different cases are obtained as for addition, with the addition operator replaced by the multiplication operator: $g(a, x) = a \cdot c$, $g(a, x) = x \cdot c$, $g(a, x) = a \cdot a$, $g(a, x) = x \cdot x$, or $g(a, x) = a \cdot x$. The first case leads to an unbounded function, the second to a re-scaled Gaussian function. The third $g(a, x) = a \cdot a$, leads to $\frac{d\mu}{dx} = \mu^3$ and hence to $\frac{d\mu}{\mu^3} = dx$. Integrating, we obtain $-\frac{1}{2\mu^2(x)} = x + C$, i.e., $\mu(x) = \sqrt{-2(x + C)}$. This expression is

not defined for large positive x, so it should also be dismissed. The fourth case $g(a,x) = x \cdot x$, leads to $\frac{d\mu(x)}{dx} = \mu(x) \cdot x^2$ hence $\frac{d\mu(x)}{\mu(x)} = x^2 \cdot dx$. Integrating, obtains $\ln(\mu(x)) = \frac{1}{3} \cdot x^3 + C$, and hence $\mu(x) = \exp\left(\frac{1}{3} \cdot x^3 + C\right)$, which is not bounded, so it has to be dismissed.

Finally, the last case, $g(a,x) = a \cdot x$, yields $\frac{d\mu(x)}{dx} = \mu(x)^2 \cdot x$ hence $\frac{d\mu(x)}{\mu^2(x)} = x \cdot dx$. Integrating, we obtain $-\frac{1}{\mu(x)} = \frac{1}{2} \cdot x^2 + C$, hence $\mu(x) = -\frac{1}{\frac{1}{2} \cdot x^2 + C}$. This is not a membership function, but adding unary minus, i.e., by considering $g(a,x) = -a \cdot x$, leads to $\mu(x) = \frac{1}{\frac{1}{2} \cdot x^2 + C}$, i.e., what we called a Cauchy membership function.

Resulting membership functions: A membership function μ is said to be *normal* if $\sup_x \mu(x) = 1$. Normal membership functions are preferred as they are considered to represent a fully defined concept. It is easy to see that $\sup_x 1/(x^2/2 + C) = 1/C$, so if this were to be a normal membership function, C must be equal to 1. Thus, the resulting membership function is shown in equation (3).

$$\mu(x) = \frac{1}{1 + \frac{x^2}{2}}. \tag{3}$$

Taking into account that the numerical value of a physical quantity depends on the choice of the measuring unit and on the choice of the starting point, changing a measuring unit and/or a starting point, a new numerical values X can be obtained from previous values x by a linear transformation $X = k \cdot x + a$, where k is the ratio of the measuring units and a is the difference in starting points. (A classical example is the relation between temperature t_C in Celsius and temperature t_F in Fahrenheit: $t_F = 1.8 \cdot t_C + 32$.)

When the original values x are described by the membership function (3), then, the membership function for X is obtained by substituting in (3), the expression $x = \frac{X-a}{k}$. This leads to the membership function shown in the equation (4),

$$\mu_X(X) = \frac{1}{1 + \frac{(X-a)^2}{2k^2}}. \tag{4}$$

or, by letting $2k^2 = K^2$ to the Cauchy membership function shown in equation (5).

$$\mu_{cauchy}(X) = \frac{1}{1 + \left(\frac{X-a}{K}\right)^2}. \tag{5}$$

Similarly, substituting $x = \frac{X-a}{k}$ into the expression $\mu(x) = \exp\left(-\frac{x^2}{2}\right)$, we obtain

$$\mu_X(X) = \exp\left[-\frac{(X-a)^2}{2k^2}\right], \tag{6}$$

or, in terms of the new parameter K:

$$\mu_{gaussian}(X) = \exp\left[-\left(\frac{X-a}{K}\right)^2\right]. \tag{7}$$

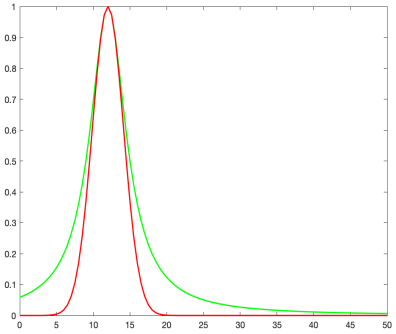

Figure 1: The Cauchy and Gaussian membership functions for $a = 12$, $K = 3$.

FIGURE 1 for μ_C and μ_G, for the same choice of the parameters a and K, shows that these functions are quite similar.

It is easy to see that in each case, the respective derivatives, $\mu'(x)$ can be easily calculated in terms of $\mu(x)$ and the derivative of $\left(\frac{X-a}{K}\right)^2$. Indeed, for the Cauchy membership function, its derivative is as given in the equation (8), while for the Gaussian function as shown in the equation (9).

$$\mu'_{cauchy}(X) = -2\mu_C^2(X)\frac{X-a}{K} \tag{8}$$

$$\mu'_{gaussian}(X) = -2\mu_G(X)\frac{X-a}{K} \tag{9}$$

On what basis should we prefer one of these membership functions? Recall that selecting a differentiable membership problem, and moreover a function whose derivative can be computed in terms of the function itself, results in an efficient calculation of the gradient of the cost function associated to a learning task. As previously mentioned, using fuzzy sets presents the additional advantage of being able to use fuzzy logic in reasoning tasks. When using the standard fuzzy logic operators for conjunction and disjunction (respectively $\min$ and $\max$), Cauchy and Gaussian membership functions produce the same result. However, for the standard negation operator ($\overline{\mu}(\cdot) = 1 - \mu(\cdot)$), the results are different as it can be seen in the equations (10) and (11) below.

$$\begin{aligned}\overline{\mu}_{Cauchy}(x) &= 1 - \mu_{Cauchy}(x) = 1 - \frac{1}{1+\frac{(x-a)^2}{K^2}} = \frac{1+\frac{(x-a)^2}{K^2}-1}{1+\frac{(x-a)^2}{K^2}} \\ &= \frac{\frac{(x-a)^2}{K^2}}{1+\frac{(x-a)^2}{K^2}} = \frac{(x-a)^2}{K^2+(x-a)^2}\end{aligned} \tag{10}$$

$$\overline{\mu}_{Gaussian}(x) = 1 - \mu_{Gaussian}(x) = 1 - e^{-\frac{(x-a)^2}{K^2}}. \tag{11}$$

This means that for the same tuple of parameters (a, K), the Cauchy membership function is slightly fuzzier than the Gaussian membership and therefore better able to distinguish between data points represented by the fuzzy set, than the Gaussian membership function. It is also expected that the use of the Cauchy membership function in computing the gradient of the cost function is more efficient.

3. CONCLUSION

This paper considered the problem of choosing a membership function from a theoretical - axiomatic-like perspective - of efficiency of learning. It was found that given the criteria for the *simplest fuzzy learning*, the membership functions which satisfy these criteria are the Cauchy and Gaussian membership functions shown in the equations (5) and (7) respectively, with the Cauchy function to be preferred.

4. ACKNOWLEDGMENTS

This work was supported in part by a grant from the "la Caixa" Banking Foundation (ID 100010434), whose code is LCF / BQ / AA19 / 11720045, by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology, and by the NSF CBET grant 1936908. It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

References

[1] Zadeh LA. "Fuzzy sets". *Information and Control*. 1965; 8:338–353.

[2] Zadeh LA. The Concept of a Linguistic Variable and Its Application to Approximate Reasoning—I.. Information sciences. 1975;8:199–249.

[3] Belohlavek R, Dauben JW, Klir GJ. Fuzzy Logic and Mathematics: A Historical Perspective. Oxford University Press, New York, 2017.

[4] Klir G, Yuan B. Fuzzy Sets and Fuzzy Logic, Theory and Applications. Prentice Hall, Upper Saddle River, New Jersey, 1995.

[5] Mendel JM. Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions. Springer, Cham, Switzerland, 2017.

[6] Nguyen HT, Walker CL, Walker EA. A First Course in Fuzzy Logic. Chapman and Hall/CRC, Boca Raton, Florida. 2019.

[7] Novák V, Perfilieva I, Močkoř J. Mathematical Principles of Fuzzy Logic. Kluwer, Boston, Dordrecht. 1999.

[8] Visa S, Ralescu A. Data-Driven Fuzzy Sets for Classification. International Journal of Advanced Intelligence Paradigms. 2008; 1:3–30.

[9] Ralescu A, Visa S. Obtaining Fuzzy Sets Using Mass Assignment Theory-Consistency With Interpolation. In NAFIPS 2007-2007,Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society. IEEE. 2007: 436–440.

[10] Viaña J, Cohen K. “Fuzzy-Based, Noise-Resilient, Explainable Algorithm for Regression”, Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society NAFIPS’2021. West Lafayette, Indiana. 2021.

[11] Viaña J, Ralescu S, Cohen K, Ralescu A, Kreinovich V. “Extension to Multi-Dimensional Problems of a Fuzzy-Based Explainable and Noise-Resilient Algorithm”, Proceedings of the 14th International Workshop on Constraint Programming and Decision Making CoProd’2021, Szeged, Hungary. 2021.

[12] Sheskin D J. Handbook of Parametric and Non-Parametric Statistical Procedures. Chapman & Hall/CRC, London, UK. 2011.