Learning Membership Functions for Fuzzy Sets through Modified Support Vector Clustering

Dario Malchiodi[1] and Witold Pedrycz[2]

[1] Università degli Studi di Milano, Italy
malchiodi@di.unimi.it
[2] University of Alberta, Canada
wpedrycz@ualberta.ca

Abstract. We propose an algorithm for inferring membership functions of fuzzy sets by exploiting a procedure originated in the realm of support vector clustering. The available data set consists of points associated with a quantitative evaluation of their membership degree to a fuzzy set. The data are clustered in order to form a core gathering all points definitely belonging to the set. This core is subsequently refined into a membership function. The method is analyzed and applied to several real-world data sets.

1 Introduction

Designing fuzzy sets has been one of the pivotal problems in the methodology and practice of the technology of fuzzy sets. Fuzzy sets come with different interpretations, cf. [1]. There are several general approaches ranging from expert-driven methods to data-driven techniques and an entire spectrum of hybrid-like strategies combining these two development modes, cf. [2]. Various shapes of membership functions are proposed [3], sometimes being directly linked with the ensuing computational facets of fuzzy sets; here we can refer to triangular fuzzy sets and their role in fuzzy modeling and a degranulation process [2,4]. Intensive pursuits in the construction of membership functions are not surprising at all: evidently fuzzy sets form a backbone of fuzzy models, fuzzy classifiers and fuzzy reasoning schemes. Fuzzy sets used in these constructs directly impact their performance as well as contribute to the interpretability (readability) of these modeling constructs. Fuzzy sets formed through an expert-driven approach are reflective of the perception of concepts captured by humans; however the estimation process could exhibit some inconsistencies associated with the elicitation process itself (bottleneck of knowledge acquisition). On the other hand, data-driven approaches rely on available experimental data and fuzzy sets obtained in this manner are reflective of the nature of the available experimental evidence (which is going to be used intensively when forming fuzzy predictors or classifiers). In this domain, we encounter techniques using which fuzzy sets (treated as information granules) arise as a summarization of numeric data; one can refer here to fuzzy clustering or other mechanisms of vector quantization [5]. With this regard a prudent formulation of the optimization process and its relevance *vis-à-vis* the semantics of fuzzy set(s) to be developed is of paramount relevance.

F. Masulli, G. Pasi, and R. Yager (Eds.): WILF 2013, LNAI 8256, pp. 52–59, 2013.

Having this mind, we propose a modified support vector clustering in which we take advantage of the formulation and the nonlinear nature of the optimization problem falling within the realm of well-established methods of support vector machines. This formulation supports a construction of diversified membership functions.

A thorough parametric analysis of the resulting construct is presented. We demonstrate how the parameters (and a tradeoff of their values) of the method impact the shape (trapezoidal, quadratic, and bimodal) of membership function of the fuzzy set being formed. A series of illustrative examples is provided to visualize the flexibility of the construct considered here.

The paper is structured as follows: we start with a suitable modification of the support vector clustering algorithm and elaborate on a selection of numeric values of the essential parameters of the method. Subsequently, we present a series of experiments showing in detail on how membership functions are constructed.

2 Modifying the SV Clustering Algorithm

Let a sample $\{x_1, \ldots, x_m\}$ in a domain X be given, together with an associated set of membership grades $\{\mu_1, \ldots, \mu_m\}$ to some unknown fuzzy set A. The problem of inferring μ_A can be divided into two parts, namely: i) determining the *shape* of A, and ii) inferring the *parameters* of the membership function μ_A. These tasks are addressed by starting from the following hypothesis.

- Set $A_1 = \{x \in X \text{ s. t. } \mu_A(x) = 1\}$ contains all points in X whose images through a mapping Φ belong to a sphere of unknown center a and radius R.
- The membership $\mu_A(x)$ only depends on the distance between $\Phi(x)$ and a.

It has been shown that the set A_1 can be estimated through a modified support-vector clustering procedure [6] provided with $x_1, \ldots, x_m$ and $\mu_1, \ldots, \mu_m$: the problem is concerned with searching for the smallest sphere, having a and R respectively as center and radius, enclosing the images of $x_1, \ldots, x_m$ produced through a transformation Φ. More precisely, we use from a starting point the typical relaxation of this problem based on slack variables $\xi_1, \ldots, \xi_m$. As our target is that of learning a fuzzy set having as inputs some points $x_1, \ldots, x_m$ and their membership values $\mu_1, \ldots, \mu_m$, we consider the constraints in the form:

$$\mu_i ||\Phi(x_i) - a||^2 \leq \quad \mu_i R^2 + \xi_i \ , \tag{1}$$

$$(1 - \mu_i)||\Phi(x_i) - a||^2 \geq (1 - \mu_i)R^2 - \tau_i \ , \tag{2}$$

$$\xi_i \geq 0, \tau_i \geq 0 \ . \tag{3}$$

It is easy to see that when $\mu_i = 1$ the constraints read in the same way as those in the problem of support vector clustering. In other words, we try to confine the images of x_i through Φ within a sphere centered at a and having radius R. On the other hand, when $\mu_i = 0$, the same set of constraint model the opposite target, i.e., exclusion of $\Phi(x_i)$ from the sphere.

Thus we can consider the following extension of the support vector clustering procedure: minimize $R^2 + C\sum(\xi_i + \tau_i)$ under constraints (1-3). Its Wolfe dual formulation is concerned with the maximization of $\sum_{i=1}^{m}(\alpha_i\mu_i - \beta_i(1-\mu_i))k(x_i,x_i) - \sum_{i,j=1}^{m}(\alpha_i\mu_i - \beta_i(1-\mu_i))(\alpha_j\mu_j - \beta_j(1-\mu_j))k(x_i,x_j)$ subject to the constraints $\sum_{i=1}^{m}(\alpha_i\mu_i - \beta_i(1-\mu_i)) = 1$ and $0 \leq \alpha_i, \beta_i \leq C$, where k denotes the kernel function associated to the dot product computation in the image of Φ (that is, $k(x_i,x_j) = \Phi(x_i)\cdot\Phi(x_j)$). Denoting with a star the optimal value for a variable, Karush-Kuhn-Tucker (KKT) conditions [7] read

$$\alpha_i^* \left({R^*}^2 \mu_i + \xi_i^* - \mu_i ||\Phi(x_i) - a^*||^2 \right) = 0 \ , \tag{4}$$

$$\beta_i^* \left((1-\mu_i)||\Phi(x_i) - a^*||^2 - {R^*}^2(1-\mu_i) + \tau_i^* \right) = 0 \ , \tag{5}$$

$$\gamma_i^* \xi_i^* = 0, \quad \delta_i^* \tau_i^* = 0 \ . \tag{6}$$

It is easy to show that when either $0 < \alpha_i^* < C$ or $0 < \beta_i^* < C$ it will necessary hold both $\xi_i^* = 0$ and $||\Phi(x_i) - a^*|| = {R^*}^2$. Thus the corresponding x_i has an image through Φ lying on the border of the learnt sphere S and will be called *support vector*. KKT conditions show that:

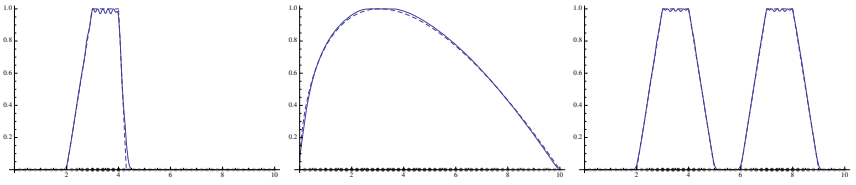
- $\alpha_i^* = 0$ implies $\xi_i^* = 0$ and $R^2(x) \leq {R^*}^2$, so $\Phi(x_i)$ lies in S or in its surface,
- $\alpha_i^* = C$ implies $R^2(x) = {R^*}^2 + \frac{\xi_i^*}{\mu_i}$, thus $\Phi(x_i)$ doesn't lie inside S,
- $\beta_i^* = 0$ implies $\tau_i^* = 0$, so that $R^2(x) \geq {R^*}^2$, thus $\Phi(x_i)$ doesn't lie inside S,
- $\beta_i^* = C$ implies $R^2(x) = {R^*}^2 - \frac{\tau_i^*}{1-\mu_i}$, thus $\Phi(x_i)$ doesn't lie outside S,

where $R^2(x) = ||\Phi(x) - a^*||^2$. Given any point $x \in X$, it can be shown that $R^2(x) = k(x,x) - 2\sum_{i=1}^{m}(\alpha_i^*\mu_i - \beta_i^*(1-\mu_i))k(x,x_i) + \sum_{i,j=1}^{m}(\alpha_i^*\mu_i - \beta_i^*(1-\mu_i))(\alpha_j^*\mu_j - \beta_j^*(1-\mu_j))k(x_i,x_j)$ so that it is possible to compute the distance between the center of the learnt sphere and the image of the given point x. In particular, all points x with membership $\mu_A(x) = 1$ satisfy $R^2(x) \leq R_1^2$, where $R_1^2 = R^2(x_i)$ for any support vector x_i. Moreover, as R^2 spans between a minimum and a maximum value when the membership value of its argument decreases from 1 to 0, the membership function μ_A can then be reconstructed in the following way:

- scaling R^2 to $R'(x) = \frac{M - R^2(x)}{M - R_1^2}$, where $M = \max_x R^2(x)$, so that R' approaches 0 and 1, respectively, when R^2 approaches its maximum and R_1^2;
- approximating μ_A with the function

$$\widehat{\mu}(x) = \begin{cases} 1 & \text{if } R'(x) \geq 1 \ , \\ R'(x) & \text{otherwise} \ . \end{cases} \tag{7}$$

The proposed procedure can produce membership functions of different shape. Figure 1 shows examples of the output for three different unidimensional membership functions, namely a trapezoidal, a quadratic and a bimodal one. In all



(a) (b) (c)

Fig. 1. Output of the proposed procedure (dashed curves) for different unidimensional membership functions (plain curves), inferred from samples of 50 item each (each sample point is drawn as a bullet colored according to its membership value, ranging from gray to black)

experiments we used a sample of $m = 50$ points uniformly distributed across the universe of discourse, associated with the corresponding membership value.

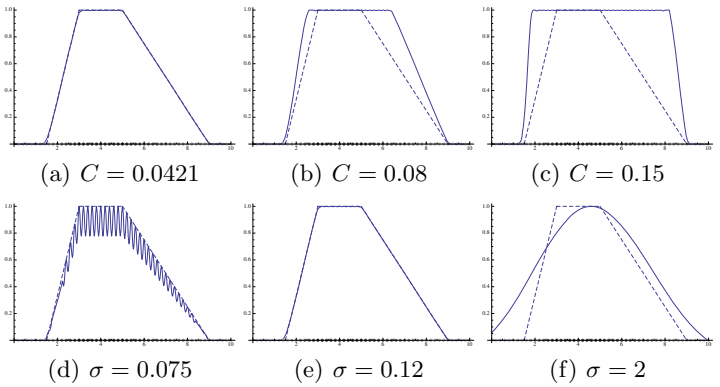
Inferring a membership function requires to strike the trade-off parameter C, as well as additional kernel parameters, an operation which is known in the literature as *model selection* [8]. In order to suitably select among the available methodologies it is worth studying the properties of parameters and their relations with the problem under study.

Figure 2 shows the results of an experiment aimed at understanding the role of involved parameters: having fixed: (i) a membership function (the dashed trapezoid in all graphs), (ii) a labeled sample, and (iii) a Gaussian kernel of parameter $\sigma = 0.12$ (see the beginning of Sect. 3), the learning procedure has been run several times using different values for C. The graphs in Fig. 2(a)–(c) highlight how an increase in C causes an enlargement of the inferred fuzzy set's *core*, intended as the subset of X whose elements are assigned unit membership. In particular, as C reaches the unit value the fuzzy set tends to a regular set enclosing all points in the labeled sample having non-zero membership values.

Similarly, we can start from the same membership function and labeled sample, set C to the best value found during the previous run, and change σ. The results, summarized in Fig. 2(d)–(f), show how the role of this parameter is that of modifying the *shape* of the membership function, which becomes more plastic as σ decreases toward zero. This experiment suggests a three-phase procedure for finding the optimal values for C and σ consisting in: 1. selecting a value C_0 in order to include in the inferred fuzzy set's core all points having unit membership; 2. selecting a value σ_0 in order to reasonably fit the data; 3. performing a fine-grained grid search centered around C_0 and σ_0.

3 Experiments

In all applications described in this paper the procedure relied on the Gaussian kernel defined by $k(x_1, x_2) = \exp\left(-||x_1 - x_2||^2/(2\sigma^2)\right)$. When using this kind of kernel [9] the optimization problem simplifies to the minimization of $\sum_{i,j=1}^{m}(\alpha_i\mu_i - \beta_i(1-\mu_i))(\alpha_j\mu_j - \beta_j(1-\mu_j))k(x_i, x_j)$; indeed, a Gaussian kernel



(a) $C = 0.0421$ (b) $C = 0.08$ (c) $C = 0.15$

(d) $\sigma = 0.075$ (e) $\sigma = 0.12$ (f) $\sigma = 2$

Fig. 2. (a)–(c): Increasing C has the effect of enlarging the learnt membership function core. (d)–(f): Increasing σ has the effect of changing the learnt membership function shape. Each graph is labelled with the corresponding parameter value.

k satisfies $k(x, x) = 1$, so that the constraints insure the equivalence between the original objective function and $1 - \sum_{i,j=1}^{m}(\alpha_i\mu_i - \beta_i(1-\mu_i))(\alpha_j\mu_j - \beta_j(1-\mu_j))k(x_i, x_j)$.

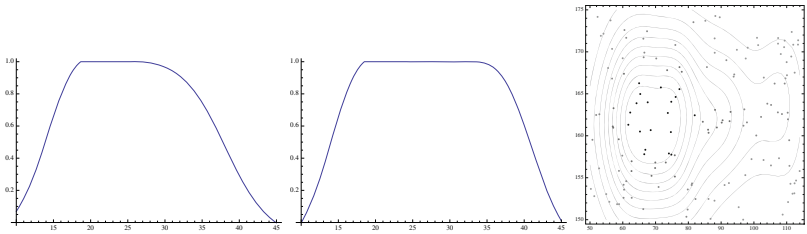
The computation of M was carried out used a Monte Carlo maximization and choosing a suitable number of samples in each experiment.

3.1 Inferring Membership Functions from Real-World Data

As a first example consider the body mass index (BMI) defined as the ratio between the weight and the squared height of a person, respectively measured in kilograms and meters. The World Health organization uses this quantity as an age- and gender-independent index for classification of weight categories in adult people, according to Table 1 [10]. Focusing on the category of *normal weight* we selected two mappings μ^1 and μ^2, shown in the table, associating each BMI range to a membership value. Subsequently we drew samples of 150 BMI values located uniformly in the interval $[10, 45]$ and computed their membership value.

Table 1. Classification of weight in function of the BMI, according to the World health organization [10]. Columns μ^1 and μ^2 show the values giving rise to the learnt membership functions shown in Fig. 3(a) and (b), respectively.

Classification	BMI range	μ^1	μ^2	Classification	BMI range	μ^1	μ^2
Severe thinness	BMI < 16	0	0	Pre-obese	$25 \leq$ BMI < 30	0.5	0.7
Moderate thinness	$16 \leq$ BMI < 17	0.2	0.4	Obese class I	$30 \leq$ BMI < 35	0.2	0.4
Mild thinness	$17 \leq$ BMI < 18.5	0.5	0.7	Obese class II	$35 \leq$ BMI < 40	0.1	0.2
Normal range	$18.5 \leq$ BMI < 25	1	1	Obese class III	BMI ≥ 40	0	0



(a) μ^1 (b) μ^2 (c)

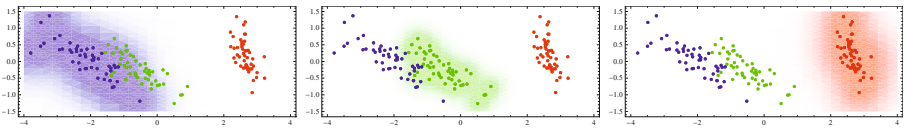
Fig. 3. (a)–(b): Learnt membership functions for normal weight according to Table 1, respectively referring to the values shown in columns μ^1 and μ^2 of the table. (c): Inferred membership function for the fuzzy set expressing the notion of normal physique in adult women in the US, in function of weight (X axis, measured in kilograms) and height (Y axis, measured in centimeters).

This allowed us to infer the membership functions (one for each mapping) shown in Fig. 3(a)–(b), setting $\sigma = 4$ and $C = 0.05$. Note how learnt membership function's shape is affected by the way categories are associated to numeric values for memberships. This is a key aspect for accommodating available domain knowledge coming from the experts in the field.

The proposed methodology is not confined to single-dimensional problems. Indeed, the kernel trick allows the inference to consider fuzzy sets defined on any space over which a kernel can be defined. Consider for instance the fuzzy notion of *normal physique* defined in terms of weight and height of a person. Figure 3(c) shows the results of a toy experiment aimed at capturing this notion, having as a starting point the distribution of weight and height, respectively measured in kilograms and centimeters, in adult women in the US [11]. Dividing the observation range in function of the data percentiles it is possible to obtain two functions μ_{weight} and μ_{height} approximating the corresponding memberships. Finally, considering a sample of 150 points uniformly drawn in $[50, 114] \times [150, 175]$ (the Cartesian product of the operational ranges in observed data) and building the membership value of each of its element (w, h) as $\mu(w, h) = \mu_{\text{weight}}(w)\mu_{\text{height}}(h)$, the proposed procedure learnt the membership function shown in Fig. 3(c).

3.2 Inferring Membership Functions in Absence of Membership Values

The method is also applicable to datasets not explicitly mentioning membership values. Consider for instance the Iris dataset [12], introduced by Fisher in 1936 and gathering 150 samples from three different species of the iris flower (namely, *Iris setosa*, *Iris virginica* and *Iris versicolor*). The observations, described through length and width of the petal and the sepal, are assigned to one of the previously mentioned species. The proposed learning procedure can be applied as follows: focusing on a given class, say *Iris setosa*, denote $\{x_1, \ldots, x_{150}\}$ the dataset observations and set $\mu_i = 1$ if x_i belongs to class *Iris setosa*, and 0



(a) ($C = 10$, $\sigma = 0.5$) (b) ($C = 100$, $\sigma = 0.225$) (c) ($C = 10$, $\sigma = 0.5$)

Fig. 4. Scatter plot of the Iris dataset and inferred membership functions for the corresponding classes. Bullets represent samples projected on their two first principal components, and colored according to their classes (in blue, green and red respectively for *Iris virginica*, *Iris versicolor*, and *Iris setosa*). Each graph also shows the density plot of the inferred membership function.

otherwise. Apply subsequently the learning procedure in order to infer a membership function μ_{setosa}. Idem for membership functions $\mu_{\text{virginica}}$ and $\mu_{\text{versicolor}}$. Given an observation x, assign it to the class it belongs to with maximal membership grade. Figure 4 shows a density plot of the membership functions inferred after application of the PCA procedure [13] selecting the first two principal components, for sake of visualization, and using a Gaussian kernel. Each plot shows the class it refers to, as well as the used values for parameters C and σ, chosen through a trial and error procedure.

We performed a more extensive experiment involving a repeated holdout scheme, in which 70% of a random permutation of the sample was used in order to infer the three membership functions, using the parameters highlighted in Fig. 4; the latter were subsequently tested on the remaining 30% of the data. Table 2 resumes average and standard deviation of the obtained error both in the training and the testing phase of 500 such procedures, starting each time from a different permutation and analyzing two, three and four principal components. These results show how even a very simple learning strategy (no complex procedures for fine tuning the parameters' choice such as a cross-validation) lead to an average test performance around 95%.

Table 2. Results of 500 holdout procedures on the Iris dataset. Each row shows average and standard deviation (columns Avg. and Stdev., respectively) of train and test error, in function of the number of principal components extracted from the original sample.

N. of principal components	Train error		Test error	
	Avg.	Stdev.	Avg.	Stdev.
2	0.00488	0.00653	0.04720	0.03143
3	0.00152	0.00349	0.06067	0.03128
4	0.00169	0.00374	0.05738	0.03347

4 Conclusions

This paper introduced a method for inferring the membership function to a fuzzy set on the basis of partial information, consisting in two finite sets: the former containing a sample of points, and the latter gathering measurements of the membership grades for points in the former set. The method relies on a special support vector clustering for the provided points, which is subsequently transformed into the inferred membership function.

References

1. Dubois, D., Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems 90, 141–150 (1997)
2. Pedrycz, W.: Granular Computing: Analysis and Design of Intelligent Systems. CRC Press/Francis Taylor, Boca Raton (2013)
3. Nguyen, H., Walker, E.: A First Course in Fuzzy Logic. Chapman Hall, CRC Press, Boca Raton (1999)
4. Pedrycz, W.: Why triangular membership functions? Fuzzy Sets & Systems 64, 21–30 (1994)
5. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Transactions on Communications COM-28(1), 84–95 (1988)
6. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support vector clustering. Journal of Machine Learning Research 2, 125–137 (2001)
7. Fletcher, R.: Practical methods of optimization, 2nd edn. Wiley-Interscience, New York (1987)
8. Guyon, I., Saffari, A., Dror, G., Cawley, G.: Model selection: Beyond the bayesian/frequentist divide. J. of Machine Learning Research 11, 61–87 (2010)
9. Evangelista, P.F., Embrechts, M.J., Szymanski, B.K.: Some properties of the gaussian kernel for one class learning. In: Marques de Sá, J., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4668, pp. 269–278. Springer, Heidelberg (2007)
10. Consultation, W.E.: Appropriate body-mass index for asian populations and its implications for policy and intervention strategies. Lancet 363(9403), 157–163 (2004)
11. McDowell, M.A., Fryar, C., Ogden, C.L., Flegal, K.M.: Anthropometric reference data for children and adults: United states, 2003–2006. National health statistics reports, vol. 10. National Center for Health Statistics, Hyattsville, MD (2008), `http://www.cdc.gov/nchs/data/nhsr/nhsr010.pdf` accessed (May 2012)
12. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2), 179–188 (1936)
13. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433–459 (2010)