Fuzzy Set Similarity using a Distance-Based Kernel on Fuzzy Sets

Jorge Guevara, Roberto Hirata, Stéphane Canu

Fuzzy Set Similarity using a Distance-Based Kernel on Fuzzy Sets

Jorge Guevara[1], Roberto Hirata Jr.[1], and Stéphane Canu[2]

[1]Departament of Computer Science, University of Sao Paulo
[2]INSA de Rouen - LITIS, Normandie Université

1 Introduction

Fuzzy sets similarity is an important topic of research due to its several theoretical and practical applications. A recent paper on similarity and dissimilarity measures between fuzzy sets [9] reviews several axiomatic definitions and plot the relationship between them. Fuzzy sets distances are also reviewed in the context of image processing and pattern recognition under imprecision by Bloch [5]. In this chapter, we present a new kind of similarity measure between fuzzy sets having a geometric interpretation in functional spaces. We will use a well-know concept from kernel methods, *the kernel*, to define a new class of similarity measures between fuzzy sets [27, 1, 4]. Kernels are real-valued functions defined on a non-empty space, and are mainly used to define similarity measures between elements in that space. Whether the kernels are positive definite or indefinite functions, always they have a geometrical interpretation either in Reproducing Kernel Hilbert Space (RKHS) or in more general spaces like Krein or pseudo-Euclidean spaces.

This work aims to show how to engineer kernels on fuzzy sets, using some well-know distances between fuzzy sets. The advantage of our approach is that is possible to have a geometrical interpretation of the similarity measure between fuzzy sets. Similarity measures between fuzzy sets computed via positive definite kernels are interpreted as inner products of two functions in a RKHS. On the other hand, more general kernels like symmetric kernels are interpreted as evaluation of functions by symmetric and bilinear forms in more general functional spaces.

Our main idea to design kernels on fuzzy sets rely on plugging a distance between fuzzy sets into the kernel definition. Kernels obtained in this way are called, *distance substitution kernels* [17]. If the distance between fuzzy sets is a metric, the resulting kernel on fuzzy sets will be positive definite, on the other hand if the distance is given by pseudo-metrics or semi-metrics, the final kernel will be a symmetric kernel.

Successfully applications of kernels on fuzzy sets will be applications where fuzzy sets are used to model imprecise or vague information, and part of the task to be solved requires the estimation of a similarity measure between fuzzy sets. Example of such applications are those within supervised and unsupervised machine learning tasks over fuzzy data. i.e. the sample contain observations given by fuzzy sets. In the same way, those kernels could be used in the analysis of heterogeneous data containing fuzzy variables. In this sense, this chapter describes an experiment on kernel hypothesis testing using heterogeneous data. This data has a subset of variables containing vague or linguistic information, that we modelled with fuzzy sets.

To conclude this section, estimating a similarity measure between fuzzy sets using kernels is important for several reasons: 1) it allow us to analyse imprecise data using fuzzy set and kernel methods. This will improve and give a new methodology for the analysis of fuzzy data in areas like data science, data mining and machine learning; 2) it can be an important tool in several areas of fuzzy mathematics, where, thanks to the theory behind kernels, fuzzy sets similarities can be interpreted from a geometrical point of view; 3) it will have several practical applications, for example, computing a similarity value between fuzzy regions of pixels in image processing, or clustering fuzzy data on data mining.

The rest of this chapter is structured as follows. We start in Section 2, giving a brief description of reproducing kernels, and showing how positive definite kernels induce a metric in a space of functions. This section also presents a class of kernels called as distance substitution kernels. Section 3 describes how to use distances between fuzzy sets to define a new class of kernels on fuzzy sets. Section 4 shows a experimental evaluation of the similarity measures presented in this chapter in a kernel hypothesis task using a real dataset. Finally, we present some conclusions in Section 4.4.

2 Reproducing kernels

Kernels are real-valued functions[1] and they are widely used as similarity measures between objects or entities [27, 1, 4]. Given two elements x and y in the set $\mathcal{X}$, a similarity measure between them can be estimated via a kernel:

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{R} && (1) \\ (x, y) &\mapsto k(x, y). && (2) \end{aligned}$$

A value $k(x, y)$ close to zero means that the two elements are not so similar. On the other hand, the higher the value $k(x, y)$ is, the more similar x is to the element y.

If k is positive definite, that is, if k satisfies the condition that:

$$\sum_{i}^{N} \sum_{j}^{N} c_i c_j k(x_i, x_j) \geq 0, \tag{3}$$

[1]There are more general definitions of kernels, but for our purposes we will only use real-valued kernels

for every possible combination of values $c_i, c_j \in \mathbb{R}$, for any elements $x_i, x_j \in \mathcal{X}$ and for all $N \in \mathbb{N}$, then the kernel evaluation has a geometric interpretation in a high dimensional Hilbert Space called as *Reproducing Kernel Hilbert Space* (RKHS). In this sense, a similarity measure between two elements in $\mathcal{X}$ computed by a positive definite kernel k is actually an inner product of two functions in a RKHS:

$$k(x, y) = \langle \phi_x, \phi_y \rangle_{\mathcal{H}}, \tag{4}$$

where ϕ_x and ϕ_y are two functions in the RKHS $\mathcal{H}$ associated to the objects x and y in the set $\mathcal{X}$. Functions ϕ_x and ϕ_y are the *representative functions* of x and y, and are mappings of the form:

$$\begin{aligned} \phi_x : \mathcal{X} &\rightarrow \mathbb{R} && (5) \\ y &\mapsto \phi_x(y) = k(x, y). && (6) \end{aligned}$$

In this sense, positive definite kernels define similarity measures for objects belonging to non-vector spaces, as is the case of the spaces given by sets of graphs, strings, logic predicates, probability measures, fuzzy sets, etc; with an geometrical interpretation in $\mathcal{H}$. For a more deeper study of kernel methods we refeer the books [27, 4].

2.1 Distance induced by positive kernels

In what follows, we will explain how positive definite kernels induce a distance in a RKHS. For that purpose, we will use the the term *distance* to denote indistinctly a metric, or a pseudo-metric, or a semi-metric.

Definition 2.1 (Metric, pseudo-metric and semi-metric)**.** A *metric* is a function: $d : \Omega \times \Omega \rightarrow \mathbb{R}_0^+$, satisfying:

1. $d(x, x') \geq 0, \forall x, x' \in \Omega$ (non negativity)
2. $d(x, x') = 0$ iff $x = x'$ (reflectivity)
3. $d(x, x') = d(x', x)$ (symmetry)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

We say that d is a *pseudo-metric* if it satisfies the axioms 1, 3, 4 and additionally satisfies $d(x, x) = 0, \forall x \in \Omega$ and $d(x, x') = 0$ for some values $x \neq x'$. If d satisfies the axioms 1, 2 and 3, it is a *semi-metric*.

In a RKHS $\mathcal{H}$ with reproducing kernel k, by virtue of (4), it is possible to compute the squared norm in $\mathcal{H}$ using kernels:

$$\|\phi_x - \phi_y\|_{\mathcal{H}}^2 = k(x, y) - 2k(x, y) + k(y, y). \tag{7}$$

It is important to notice that not only positive definite kernels could induce metrics in $\mathcal{H}$, There is another important class of kernels, the *Conditionally*

Positive Definite kernels (CPD) that are defined in the same way as Eq.3 but with the restriction of $\sum_i^N c_i = 0, \forall c_i \in \mathbb{R}$.

If k' is a CPD kernel, (by Proposition 3 in [26] and Lemma 2.1 in [3]), the following kernel is positive definite:

$$k(x,y) = \frac{1}{2}\big(k'(x,y) - k'(x,x_0) - k'(x_0,y) + k'(x_0,x_0)\big) \tag{8}$$

Now, if $\phi_x, \phi_y \in \mathcal{H}$ are induced by a kernel as the one given by (8), then (7) could be rewritten as[2]:

$$\|\phi_x - \phi_y\|^2_{\mathcal{H}} = -k'(x,y) + 1/2(k'(x,x) + k'(y,y)). \tag{9}$$

It is still possible to relax the condition of positive definiteness for kernels by only requiring symmetric kernels to be used to define distances. The geometric interpretation in that case is given by linear spaces with symmetric bilinear forms[3]. Of course, such induced distances could violate the definition of metric, nevertheless they would induce zero-diagonal symmetric distances. Symmetric and non-symmetric kernels had been used in machine learning problems with state-of-the-art results [18] [19, 21, 22]. Some work had been done to give a geometrical interpretation for indefinite kernels in pseudo-euclidean spaces [24, 18] and Krein spaces [23]. In such spaces bilinear forms are not necessarily positive definite and norms do not define metrics.

2.2 Distance substitution kernels

This class of kernels, are kernels based on the idea of using a distance inside the kernel definition [17], and are motivated by the distance based learning paradigm [24]. That is, the task of learning statistical models from data, where the observations are not given explicitly but only a distance evaluation over pairs of observations is known beforehand. Following Definition 2.1, if a metric $d(x,y)$ is defined by $\|\phi_x - \phi_y\|_{\mathcal{H}}$ using some positive definite kernel k, i.e., d is an isometric embedding onto $\mathcal{H}$, then the following expressions are positive definite kernels:

$$1)\ \langle x,y\rangle_d^{x_0}, \quad 2)\ \big(\alpha + \gamma\langle x,y\rangle_d^{x_0}\big)^{\beta}, \quad 3)\ \exp(-\gamma d(x,y)^2), \tag{10}$$

where $\alpha, \gamma \in \mathbb{R}^+$, $\beta \in \mathbb{N}$, and

$$\begin{aligned}
\langle x,y\rangle_d^{x_0} &\overset{\text{def}}{=} \big\langle(\phi_x - \phi_{x_0}),(\phi_y - \phi_{x_0})\big\rangle_{\mathcal{H}} && (11)\\
&= \frac{1}{2}\big(\|(\phi_x - \phi_{x_0})\|^2_{\mathcal{H}} + \|(\phi_y - \phi_{x_0})\|^2_{\mathcal{H}} - \|\phi_x - \phi_y\|^2_{\mathcal{H}}\big) && (12)\\
&= \frac{1}{2}\big(d(x,x_0)^2 + d(y,x_0)^2 - d(x,y)^2\big) && (13)
\end{aligned}$$

[2] See Proposition 3.2 in [3] for details.

[3] If k is symmetric, then $k(x,x') = Q(k(.,x),k(.,x'))$, where Q is a symmetric and bilinear form, with reproducing property $Q(k(.,x),f) = f(x)$. See Proposition 6 in [26] for details.

By the definition of a CPD kernel, it is possible to prove that $-d(x,y)^p, p \in [0,2]$ is a CPD kernel. Note, that $\phi_{x_0} \in \mathcal{H}$ is an arbitrary origin. Relaxing the definition for d and allowing the use of pseudo-metrics and semi-metrics, the kernels mentioned are not longer positive definite, instead they are symmetric kernels. The geometric interpretation is guaranteed in pseudo-Eucledian spaces or Krein spaces [7, 19, 21].

3 Kernels on fuzzy sets via fuzzy distances

This section describes a new class of kernels that actually are distance substitution kernels, where the distance function is a distance between fuzzy sets. That is, they are formulated by plugging a distance between fuzzy sets into the kernel definition.

3.1 Kernels on fuzzy sets

Similarity measures for fuzzy sets using kernels has two main advantages. The first one is related to the geometrical interpretation of the similarity measure in a space of functions. In this way, a positive definite kernel on fuzzy sets could induce a topology in the set of all the fuzzy sets. The another advantage is that kernels on fuzzy sets could be used in a modular fashion in kernel methods like support vector machines, kernel PCA, Gaussian process and so on. Such advantages could leverage the development of many applications in several areas of research where the observational data contain variables that are better represented by fuzzy sets or inclusively data whose observations are naturally represented by fuzzy sets.

Kernels on fuzzy sets are mappings of the form [16, 14, 15]:

$$\begin{aligned} k : \mathcal{F}(\Omega) \times \mathcal{F}(\Omega) &\rightarrow \mathbb{R} && (14) \\ (X, Y) &\mapsto k(X, Y), && (15) \end{aligned}$$

where $\mathcal{F}(\Omega)$ denotes the set of fuzzy sets on Ω, that is

$$\mathcal{F}(\Omega) = \{X \mid X : \Omega \rightarrow [0,1]\}, \tag{16}$$

We say that the value $X(x), x \in \Omega$ is the membership degree of x in the fuzzy set X.

3.1.1 Distances on fuzzy sets induced by kernels

Distances between fuzzy sets could be formulated using kernels on fuzzy sets as follows:

$$D(X, Y) = k(X, Y) - 2k(X, Y) + k(Y, Y), \tag{17}$$

we will call such a distance as a *kernel-based distance over fuzzy sets*.

An important consequence of Equation (7) in fuzzy set theory is that if the kernel on fuzzy sets is positive definite, then Equation (7) give us a way on

defining a new set of metrics over fuzzy sets. In this sense $D(X, Y)$ is interpreted as being $\|\phi_X - \phi_Y\|$ in the RKHS $\mathcal{H}$ induced by the kernel k.

More general kernels on fuzzy sets can be used to define distances on fuzzy sets. However, the resulting distances would not be metrics. Nevertheless, it is still possible to get a zero-diagonal distances on fuzzy sets with symmetric kernels.

3.2 A distance-based kernel on fuzzy sets

Going in the other direction, if a metric on fuzzy sets in the sense of Definition 2.1 is given, then a positive definite kernel on fuzzy sets could be easily constructed from it, using the distance substitution kernel concept. That is, if D is a metric, the following kernels on fuzzy sets are positive definite

$$K(X, Y) = \langle X, Y \rangle_D^{x_0} \tag{18}$$

$$K(X, Y) = \left(\alpha + \gamma \langle X, Y \rangle_D^{X_0}\right)^{\beta} \tag{19}$$

$$K(X, Y) = \exp(-\gamma D(X, Y)^2), \tag{20}$$

where $\alpha, \gamma \in \mathbb{R}^+$, $\beta \in \mathbb{N}$, and

$$\langle X, Y \rangle_D^{X_0} \overset{\text{def}}{=} \frac{1}{2}\left(D(X, X_0)^2 + D(Y, X_0)^2 - D(X, Y)^2\right) \tag{21}$$

All of those kernels will be kernels on fuzzy sets induced by distances on fuzzy sets, we will refer to them as *distance-based kernels on fuzzy sets*.

If D is either a pseudo-metrics or a semi-metrics on fuzzy sets, function D still could be used to define kernel on fuzzy sets. Of course, those kernels are not longer positive definite but a geometrical interpretation could be given in pseudo-Euclidean spaces or Krein spaces. Furthermore, non-positive kernels have been used in several applications on machine learning problems in the past few years. See for example references [2, 7, 19, 21]

Some popular distances between fuzzy sets that could induce new kernels on fuzzy sets are given in Table 1. See [5, 25, 8, 10] and references therein for details.

4 Experiments

This section presents an experimental study using the kernels presented in this work on a hypothesis testing task. The dataset used in this experiment is a real dataset of cancer prognosis and, following Section 4.2.1, some variables or features are described by fuzzy sets rather than using specific values. We used a kernel two-sample test [12], with kernels on fuzzy sets to find out whether two samples of points follow the same distribution[4].

[4] A code of the experiments could be found at `https://github.com/jorjasso/Two-sample-kernel-test-with-fuzzy-kernels`.

$D(X, X')$	Type of distance
$\left[\int_{x\in\Omega} \lvert X(x) - X'(x)\rvert^p\right]^{1/p}$	pseudo-metric, $p \in \mathbb{N}$
$\sup_{x\in\Omega} \lvert X(x) - X'(x)\rvert$	metric
$\dfrac{\sum_{x\in\Omega} \lvert X(x) - X'(x)\rvert}{\sum_{x\in\Omega} \lvert X(x) + X'(x)\rvert}$	metric, discrete case
$\lvert E(X) - E(X')\rvert$	pseudo-metric[a]
$\sum_{x\in\Omega}[D_x(X, X') + D_x(X', X)]$	semi-metric[b]
$1 - \dfrac{\sum_{x\Omega} \min[X(x), X'(x)]}{\sum_{x\Omega} \max[X(x), X'(x)]}$	semi-metric
$1 - \sum_{x\Omega} \dfrac{\min[X(x), X'(x)]}{\max[X(x), X'(x)]}$	semi-metric
$1 - \max_{x\Omega} \min[X(x), X'(x)]$	pseudo-metric
$\int_0^1 D(X_\alpha, X'_\alpha)d\alpha$	X_α is the α-cut set and D is any metric on sets
$\dfrac{\sum_{x,x'\in\Omega} d(x, x') \min[X(x), X'(x')]}{\sum_{x,x'\in\Omega} \min[X(x), X'(x')]}$	metric

[a] $E(X) = -K \sum_{x\in\Omega}[X(x) \log X(x) + (1 - X(x)) \log(1 - X(x))]$

[b] $D_x(X, X') = X(x) \log \dfrac{X(x)}{X'(x)} + (1 - X(x)) \log \dfrac{(1 - X(x))}{(1 - X'(x))}$

Table 1: Popular distaces on fuzzy sets.

4.1 Kernel hypothesis testing

Kernels had been used to perform statistical tests to decide if two samples were drawn from the same or different distributions [6, 11, 13, 12]. In this experiment we used the kernel two-sample test approach which central concept is the Maximum Mean Discrepancy (MMD) [12] defined as follows.

Definition 4.1 (MMD)**.** Let $\mathcal{F}$ be a set of functions $f : \Omega \to \mathbb{R}$. Let X and Y be two random variables on Ω distributed according to the probability measures $\mathbb{P}$ and $\mathbb{Q}$. The MMD is defined as:

$$\mathrm{MMD}[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \sup_{f \in \mathcal{F}} \big(\mathbb{E}_{\mathrm{X}\sim\mathbb{P}}[f(\mathrm{X})] - \mathbb{E}_{\mathrm{Y}\sim\mathbb{Q}}[f(\mathrm{Y})]\big), \tag{22}$$

where $\mathbb{E}_{\mathrm{X}\sim\mathbb{P}}$ is the expectation of the random variable X distributed according to $\mathbb{P}$.

For practical purposes, it is better to compute the square of the MMD in a unit sphere of a RKHS $\mathcal{H}$ with kernel k. For a detailed discussion about the derivation of the MMD in a RKHS see [12].

$$\mathrm{MMD}^2[\mathcal{F}, \mathbb{P}, \mathbb{Q}] \quad = \quad \big\|\mathbb{E}_{\mathrm{X}\sim\mathbb{P}}[\phi_{\mathrm{X}}] - \mathbb{E}_{\mathrm{Y}\sim\mathbb{Q}}[\phi_{\mathrm{Y}}]\big\|^2_{\mathcal{H}}.$$

Given the samples $s_X = \{x_1, \ldots, x_m\}$ e $s_Y = \{y_1, \ldots, y_n\}$ distributed according to $\mathbb{P}$ and $\mathbb{Q}$, respectively. An unbiased empirical estimator is given by (Lemma 6 from [12]):

$$\begin{aligned} \mathrm{MMD}^2_u[\mathcal{F}, s_X, s_Y] \quad = \quad & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j\neq i}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \\ & + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} k(y_i, y_j) \end{aligned} \tag{23}$$

Using the samples s_X and s_Y defined above, it is possible to use the MMD to test whether the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ or the alternative hypothesis: $H_A : \mathbb{P} \neq \mathbb{Q}$ hold. To achieve that, the evaluation of the metric MMD is compared against a threshold ϵ. That is, if $\mathrm{MMD}[\mathcal{F}, s_X, s_Y] > \epsilon$, it is possible to say that the null hypothesis is rejected. Otherwise, there is not evidence to reject the null hypothesis. A Type I error occurs when H_0 is rejected in spite of the null hypothesis being true. A Type II error occurs when the null hypothesis is not rejected in spite of the distributions being different. The level of the test denoted by α is an upper bound of the Type I error and, of course, it can be used to estimate the value of the threshold ϵ. There is several methods to achive that, see Section 4 and 5 from [12] for an example. In our experiments, we estimated the threshold using bootstrap.

4.2 Two-sample test with kernels on fuzzy sets

In this section, we first describe the dataset used in the experiment and we show that some variables of the dataset contain imprecise values given by linguistic terms. Later, we show a fuzzification methodology to model the imprecise values of such variables and we finish reporting some comparative results on fuzzy sets against classical kernels.

4.2.1 Breast cancer dataset

The *Breast Cancer* dataset [20][5], provided by the Institute of Oncology at the University Medical Center in Ljubljana, Yugoslavia, has information of 286 patients, divided in two categories: 201 patients with no recurrence events of breast cancer in the first five years after the surgery, and 85 patients that show recurrence events of breast cancer in the same period of time.

The dataset has ten variables shown in Table 2. Note that the variables *age, menopause, tumor-size* and *inv-node* contain imprecise values.

Variable	Values
class	no-recurrence-events, recurrence-events
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
menopause	lt40, ge40, premeno.
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
node-caps	yes, no.
deg-malig	1, 2, 3.
breast	left, right.
breast-quad	left-up, left-low, right-up, right-low, central.
irradiat	yes, no.

Table 2: Variables of the Breast Cancer Dataset.

4.2.2 Fuzzification of the variables of the dataset

Table 2 shows that the dataset is heterogeneous, because it contains variables with precise and non-precise values. For the later kind of variables, i.e., the variables *age, menopause, tumor-size*, and *inv-nodes*, we used fuzzy sets to model or encode the imprecise values. In what follows, we will explain how the fuzzification process of those variables was performed with imprecise values.

The variable *age* was modeled with fuzzy sets using trapezoidal membership

[5]Available from `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer`

functions given by:

$$X(x;a,b,c,d) = \begin{cases} 0, & x \leq a \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \dfrac{d-x}{d-c}, & c \leq x \leq d \\ 0, & d \leq x \end{cases} \tag{24}$$

After analyzing the range of values of this variable, we used the following rule to define the parameters a, b, c, d. Let l and r be the leftmost and the rightmost values of each interval (see 2), then $a = l - 5$, $b = l$, $c = r$, $d = r + 5$. The reason behind this choice is to allow some overlapping by the membership functions. Figure 1 show the trapezoidal fuzzy sets used to represent the values of the variable *age*.

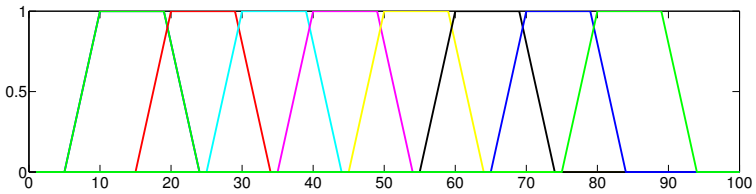

Figure 1: Fuzzification of variable *age* with trapezoidal fuzzy sets.

In the case of the variable *menopause*, we used a fuzzy set with a *Z-shaped* membership function to represent the linguistic term *lq40* (less than forty years old), that is:

$$X(x;a,b) = \begin{cases} 1, & x \leq a \\ 1 - 2\left(\dfrac{x-a}{b-a}\right)^2, & a \leq x \leq \dfrac{a+b}{2} \\ 2\left(\dfrac{x-a}{b-a}\right)^2, & \dfrac{a+b}{2} \leq x \leq b \\ 0, & x \geq b, \end{cases} \tag{25}$$

where $a = 40$ and $b = 45$. Figure 2 shows this fuzzy set as a red curve. We used a fuzzy set with a *S-shaped* membership function to model the linguistic value *ge40* (greater than forty years old), that is:

$$X(x;a,b) = \begin{cases} 0, & x \leq a \\ 2\left(\dfrac{x-a}{b-a}\right)^2, & a \leq x \leq \dfrac{a+b}{2} \\ 1 - 2\left(\dfrac{x-a}{b-a}\right)^2, & \dfrac{a+b}{2} \leq x \leq b \\ 1, & x \geq b, \end{cases} \tag{26}$$

where $a = 35$ and $b = 40$. Figure 2 shows this fuzzy set as a blue curve. To represent the linguistic value *premeno*, we used a fuzzy set with a Gaussian

membership function, that is:

$$X(x;\sigma,c) = \exp\Big(-\frac{(x-c)^2}{2\sigma^2}\Big) \tag{27}$$

where $c = 45$ and σ was estimated to be $(50-40)/(2\sqrt{2\log 2})$ [6]. Figure 2 show this fuzzy set as a yellow curve.

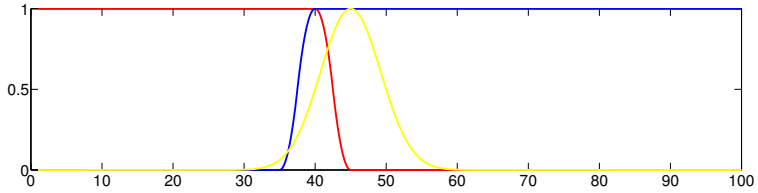

Figure 2: Fuzzification of the variable *menopause* with Gaussian (yellow), S-shaped (blue) and Z-shaped (red) fuzzy sets.

The variable *tumor-size* is modeled by three kinds of fuzzy sets. We used a *Z-shaped* fuzzy set with parameters $a = 0$ and $b = 9$ and a *S-shaped* fuzzy set with parameters $a = 50$ and $b = 59$ to model the extreme values of the variable. Several Gaussian fuzzy sets with parameters $c = (l+r)/2$ and $\sigma = (r-l)/(2\sqrt{2\log 2})$ were also used to model the other values. Figure 3 shows the fuzzy sets used to model this variable.

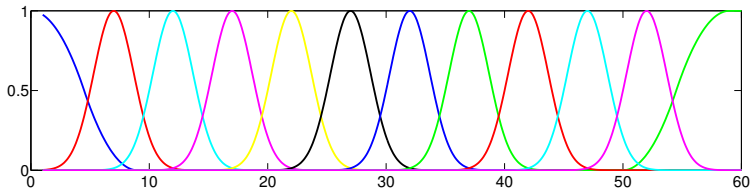

Figure 3: Fuzzification of the variable *tumor-size*, with S-shaped (leftmost), Z-shaped (rightmost) and Gaussian fuzzy sets.

Finally, we modeled the variable *inv-nodo* in the same way as the variable *tumor-size*. We used a *Z-shaped* fuzzy set with parameters $a = 0$ and $b = 5$, a *S-shaped* fuzzy set with parameters $a = 33$ and $b = 39$, and several Gaussian fuzzy sets with parameters $c = (l+r)/2$ and $\sigma = (r-l)/(2\sqrt{2\log 2})$. Figure 4 shows the fuzzy sets used to model this variable.

The other variables of this dataset are categorical variables. Table 3 shows those variables with their respective sets of values. It is also shown the discrete values that we used to represent such values.

For experimental purposes, we also present a crisp version of this dataset. To do that, we take the mean value of each interval for the case of the variables

[6] This value is known as *full width at half maximum*.

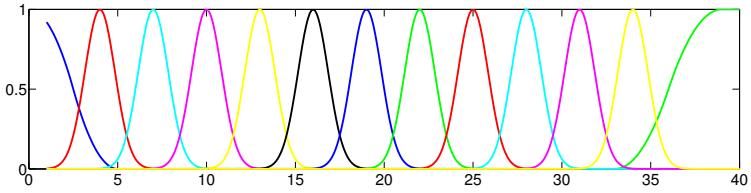

Figure 4: Fuzzification of the variable *inv-nodes* with with S-shaped (leftmost), Z-shaped (rightmost) and Gaussian fuzzy sets.

Variable	Values	Discrete values
class	no-recurrence-events	1
	recurrence-events	-1
node-caps	yes	1
	no	-1
deg-malig	1	−1
	2	0
	3	1
breast	left	1
	right	-1
breast-quad	left-up	−1
	left-low	-0.5
	right-up	0
	right-low	0.5
	central	1
irradiat	yes	1
	no	-1

Table 3: Categorical variables and their values *node-caps, deg-malig, breast* e *irradiat*.

age, tumor-size and *inv-nodes*. We set the variable *menopause* to have one of the values $\{-1, 0, 1\}$ to represent its linguistic values. The values of the other variables are the same of Table 3.

4.2.3 Experimental setting

As stated before, MMD can be used to determine whether two samples were drawn from the same distribution (null hypothesis) or not. In order to use the MMD with kernels on fuzzy sets over the fuzzy version of the breast cancer data we performed two experiments.

Experiment I: The purpose of this experiment is to verify if the MMD is able to correctly identify if a pair of samples were drawn from two *different* distributions. We set the level of the test α to the value 0.05. We applied the test one hundred times over pairs of samples, each of them having 25 observations drawn from two different distributions: patients showing breast cancer recurrence or patients without breast cancer recurrence. For each time the test is applied, we count if the samples come from the same distribution by counting the number of times the null hypothesis is rejected. In order to get the mean and variance, we performed the above procedure 250 times.

Experiment II: In this experiment, we used the same procedure as before but this time each pair of samples were chosen from the *same* category (recorrence, or non recorrence). The category whereby the samples were chosen was randomly selected. Therefore, we estimated the number the times the test is able to correctly identify if the observations were drawn from the same categories.

For comparison purposes, in addition of the fuzzified dataset, we hold a crisp version of the dataset (Section 4.2.2).

For the crisp dataset we used the linear kernel: $k(x, y) = x^T y,\ x, y \in \mathbb{R}^D$, and we denoted this kernel by k_{lin}. We also used the RBF kernel: $k(x, y) = \exp(-\gamma\|x - y\|^2),\ x, y \in \mathbb{R}^D,\ \gamma \in \mathbb{R}^+$, denoted by k_{RBF}. The γ parameter was estimated using the median heuristic, that is, we compute the median of the all possible distances between observations in the dataset and we set γ to be the inverse of this value.

In the case of the fuzzified dataset, as not all the variables of the fuzzy dataset are fuzzy sets, (see Table 2) we decided to use a convex combination of kernels:

$$k(x, y) = \sum_{i=1}^{K} \beta_i k_i(x_i, y_i), \tag{28}$$

where $\beta_i \geq 0$ and $\sum_i \beta_i = 1$. The advantage of this strategy is that different kernels can be used in different subsets of variables containing values from different domains. The resulting kernel is positive definite because any linear combination of kernels is also a positive definite kernel. Another advantage is that it is possible to optimize the values β_i from data. Using the above approach, we set the following kernels for the fuzzy dataset. We denoted by $k_\cap + k_{lin}$, the kernel resulting from the convex combination of linear kernels on crisp variables and the intersection kernel on fuzzy sets over the fuzzy variables *age, menopause,*

tumor-size and *inv-nodes*[14]. For the intersection kernel on fuzzy sets, we used the minimum T-norm operator. The resulting kernel was used in the MMD. Notice that this kernel has not bandwidth parameter, then is not necessary to perform a search to choose some optimal parameter.

Another convex combination of kernels, denoted by $k_\cap + k_{RBF}$, was given by using the RBF kernel for the crisp features and the intersection kernel on fuzzy set for the fuzzy variables. The RBF kernel parameter was given by the median heuristic but only considering the crisp dimensions.

Finally we used a convex combination of a RBF kernel for the crisp variables and a distance substitution kernel on fuzzy sets with the following metric on fuzzy sets

$$D(X, X') = \frac{\sum_{x\in\Omega} |X(x) - X'(x)|}{\sum_{x\in\Omega} |X(x) + X'(x)|}, \tag{29}$$

we used this metric on fuzzy sets to get the following kernel on fuzzy sets:

$$K_D(X, X') = \exp(-\lambda D(X, X')^2) \tag{30}$$

The kernel parameter was computed by the median heuristic over the crisp versions of the fuzzy variables. We denoted this kernel by $K_D + k_{RBF}$.

4.3 Results

Figure 5 shows a plot of Type II error against the level of the test α. Notice that the kernel $k_D + k_{RBF}$ has lower Type II error than the other ones, for several choices of α.

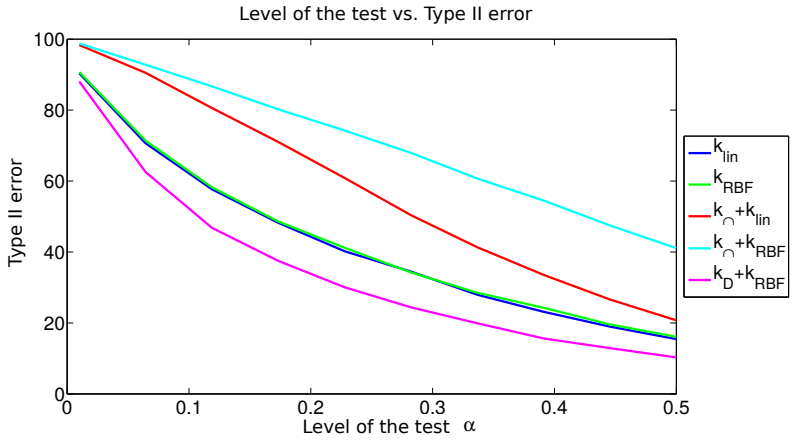

Figure 5: Level of the test α vs. Type II error using a MMD with five different kernels.

Table 4 shows the results for $\alpha = 0.05$. The first column shows the type of the experiment. The second column shows the condition of the samples. In the

third column the word *accepted* must be understood as *there is not evidence to reject the null hypothesis*, and the word *rejected* must be understood as *the null hypothesis is rejected*. The other columns show the results for the kernels mentioned in Section 4.2.3. The first row of the table shows if there is not evidence to reject the null hypothesis, that is, if the two samples were drawn from the same distribution (the same category), the test correctly says that it is the case. We observed that all the kernels perform pretty well. The second row shows the case when the null hypothesis is rejected despite of being true. That is called a Type I error. We observed that all the test have a lower Type I error, because this value is bounded by the level of the test.

We saw in the third row the case where the test wrongly says that both samples were drawn from the same distribution (there is not evidence to reject the null hypothesis) in spite of they were draw from two different ones. That is the test says that the alternative hypothesis is false despite it is true. This kind of error is called a Type II error, as the Type I error is bounded by the level of the test, the Type II error give us a way to measure the relevance of the kernels in the hypothesis testing task. We observed that the kernel on fuzzy sets $K_D + k_{RBF}$ is the best of all of them achiving lower Type II error for this dataset. One possible explanation for this is that this kernel exploit the imprecise information given by the fuzzy variables better than the others, in the sense that gives a more suited similarity measure for this kind of observations.

The fourth row of the table shows the case when the test say that the alternative hypothesis is true, when it is really true. That is the null hypothesis is rejected. We saw that kernels $k_\cap + k_{lin}$ and $k_\cap + k_{RBF}$ performs badly and the kernel $K_D + k_{RBF}$ has a highly value than the others. To conclude, we also report in Figures 6 and 7, the Type I and II errors in terms of boxplots for $\alpha = 0.05$.

Exp.	Condition	H_0	k_{lin}	k_{RBF}	$k_\cap + k_{lin}$	$k_\cap + k_{RBF}$	$k_D + k_{RBF}$
II	Same	accepted	97.63	97.98	97.38	97.51	97.92
II	Same	rejected	02.37	02.02	02.62	02.49	02.08
I	Different	accepted	73.79	73.68	92.78	93.85	**61.09**
I	Different	rejected	26.21	26.32	07.22	06.15	38.91

Table 4: Results for $\alpha = 0.05$. The null hypothesis is denoted by H_0 .

4.4 Conclusions

Similarity measures between fuzzy sets are very important because its several applications. In this chapter we proposed a new kind of similarity measure based on the idea of plugging fuzzy distances into kernel definitions. The advantage of this approach is that if the distance is a metric then the resulting similarity measure between fuzzy sets has a geometric interpretation in a RKHS, therefore, it is possible to use kernels to analyze datasets whose variables are better modeled

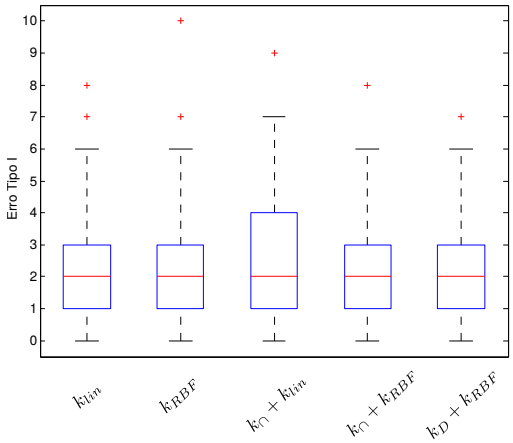

Figure 6: Type I error.

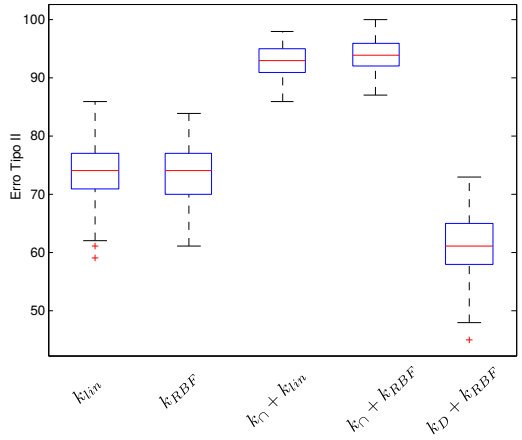

Figure 7: Type II error

using fuzzy sets. In the case of more general distances like pseudo-metrics and semi-metrics, it is still possible to have a geometric interpretation in Krein or pseudo-Euclidean spaces. The resulting kernels will be symmetric kernels, nevertheless literature shows that those kernels are still valid in machine learning tasks, given good experimental results. We also showed an experimental evaluation of kernels on fuzzy sets in the hypothesis testing task in a real dataset. As we noticed, it was possible to get a better performance using a linear combination of fuzzy kernels and the other ones for this task. We believed that this happened because fuzzy sets encodes better the imprecise information provided by some variables of this dataset, and kernels on fuzzy set are better suited to find out a good similarity measure between this fuzzy sets. Finally, we think that it is possible to apply this methodology to solve another practical problems where variables or observations are given by fuzzy sets and it is necessary to use a similarity measure to cope with the problem.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 1950.

[2] C. Bahlmann, B. Haasdonk, and H. Burkhardt. Online handwriting recognition with support vector machines - a kernel approach. In *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, pages 49–54, 2002.

[3] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, Berlin, 1984.

[4] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Kluwer Academic Boston, 2004.

[5] Isabelle Bloch. On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition*, 32(11):1873–1895, 1999.

[6] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[7] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999.

[8] BB Chaudhur and Azriel Rosenfeld. On a metric distance between fuzzy sets. *Pattern Recognition Letters*, 17(11):1157–1160, 1996.

[9] I. Couso, L. Garrido, and L. Sánchez. Similarity and dissimilarity measures between fuzzy sets: A formal relational study. *Information Sciences*, 229:122–141, 2013.

[10] Phil Diamond, Peter E Kloeden, Peter Eris Kloeden, Australia Mathematician, and Peter Eris Kloeden. *Metric spaces of fuzzy sets: theory and applications*. World Scientific, 1994.

[11] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2006.

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[13] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[14] J. Guevara, R. Hirata, and S. Canu. Positive definite kernel functions on fuzzy sets. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 439–446, July 2014.

[15] Jorge Guevara. *Modelos de apredizado supervisionado usando métodos kernel, conjuntos fuzzy e medidas de probabilidade*. PhD thesis, Universidade de Sao Paulo, Brazil, 2015.

[16] Jorge Guevara, Roberto Hirata, and Stephane Canu. Kernel functions in takagi-sugeno-kang fuzzy system with nonsingleton fuzzy input. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–8, 2013.

[17] B. Haasdonk and C. Bahlmann. Learning with distance substitution kernels. *Pattern Recognition*, pages 220–227, 2004.

[18] Bernard Haasdonk. Feature space interpretation of svms with indefinite kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(4):482–492, 2005.

[19] Bernard Haasdonk and Daniel Keysers. Tangent distance kernels for support vector machines. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 864–868. IEEE, 2002.

[20] M. Lichman. UCI machine learning repository, 2013.

[21] Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, page None, 2003.

[22] Hiroshi Shimodaira Ken-ichi Noma. Dynamic time-alignment kernel in support vector machine. *Advances in neural information processing systems*, 14:921, 2002.

[23] C.S. Ong, X. Mary, S. Canu, and A.J. Smola. Learning with non-positive kernels. In *Proceedings of the twenty-first international conference on Machine learning*, page 81. ACM, 2004.

[24] Elzbieta Pekalska, Pavel Paclik, and Robert PW Duin. A generalized kernel approach to dissimilarity-based classification. *The Journal of Machine Learning Research*, 2:175–211, 2002.

[25] Azriel Rosenfeld. Distances between fuzzy sets. *Pattern Recognition Letters*, 3(4):229–233, 1985.

[26] Bernhard Scholkopf. The kernel trick for distances. *Advances in neural information processing systems*, pages 301–307, 2001.

[27] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.