Supervised machine learning with kernel embeddings of fuzzy sets and probability measures

Jorge Guevara

Supervised machine learning with kernel embeddings of fuzzy sets and probability measures

Jorge Guevara Diaz*

Department of Computer Science,
University of Sao Paulo (USP),
05508-090, Sao Paulo, SP, Brazil.
Email: jorjasso@vision.ime.usp.br

(Summary Ph.D Thesis)

***Abstract*—This document is a summary of my Ph.D. Thesis. I was supervised by Prof. Dr. Roberto Hirata Jr. at the eScience laboratory, University of Sao Paulo, Brazil. Further, I did an one-year internship at LITIS laboratory, INSA-Rouen, University of Normandy, France, under advise of Prof. Dr. Stephane Canu. My thesis work was financed by the funding agencies: CAPES, CNPq, FAPESP grant 2011/50761-2, NAP eScience - PRP - USP from Brazil, and LITIS laboratory from France.**

I. Introduction

This research explores the problem of extending machine learning (ML) methods, specifically supervised methods, to the case of datasets given by the aggregation of several sets, i.e., datasets of the form:

$$\mathcal{T} = \{s_i\}_{i=1}^{N}, \tag{1}$$

where $N \in \mathbb{N}$ is the number of *observations* and each observation s_i is a non-empty set. Examples on methods and datasets for this task are in references: [1]–[14].

The ML pipeline for such datasets usually involves a preprocessing step that computes a feature value from each observation, with the attempt of encoding the behavior of points within the set. However, useful information may be lost, resulting in models with poor performance.

The importance of the study of this problem relies on the development of ML models for such data. Therefore, proposed solutions could improve performance, interpretability, and hence robustness in ML applications. However, a main difficulty to be solved is how to include all the information provided by the data into ML models.

Even though this problem could be solved from several points of views. i.e., neural networks, generative models, etc, we used an approach based on discriminative kernel methods, fuzzy sets theory, and probability measures. We justify this choice in the next paragraphs.

Kernel methods. They are a class of models and algorithms to perform ML on data [15]–[18]. They are based in two main concepts: 1) a similarity measure between observations called *kernel* [19]–[22]; and 2) a ML algorithm working with kernels. Indeed, a kernel is a mapping from the data space to a special kind of geometric space, for instance, a *Reproducing Kernel Hilbert Space* (RKHS), a Krein space [23] or a pseudo-Euclidean space [24]. Kernels enable to carry out operations on those high-dimensional spaces via the *kernel trick* [16]: a kernel evaluation of two points from the original space correspond to a bilinear form evaluation of two functions[1]. Therefore, a ML algorithm is implicitly defined, via the kernel trick, on those spaces. Kernels does not have constraints about the input space. Hence, there exists kernels defined on strings, graphs, images, distributions, sets, logic predicates, etc. An advantage of kernels methods is their modularity: it is possible to change the kernel definition without changing the algorithm. Examples of kernel methods are: support vector machines [25]–[27], support vector data description [28], Kernel PCA [16], Gaussian process [29], multiple kernel learnig [30].

Probability measures. They are functions used to *measure* sets. They are used to model uncertainty from a *random* perspective. i.e., uncertainty arise by chance and it is modeled from a probabilistic point of view. In this sense, we use probability measures to say that a observation s_i is a sample following a probabilistic law defined by a probability measure. Therefore, the analysis of the datasets of interest given by Equation (1), depends on the local distribution of each observation s_i. Examples of such observations, are: a set of features from a image [31]; temporal-space features [32]; object invariance features [10]; among others [1]. Probability measures are widely used in measure theory and probability theory. They are widely used within the ML community as a tool to describe and model random uncertainty in data and algorithms.

Fuzzy sets. A fuzzy set [33] is a relaxed version of a set: its indicator function, called as *membership function*, takes values within an interval. Fuzzy sets have been widely used to model *uncertanty*. In this research we used fuzzy sets to model uncertainty from a observation s_i using an *ontic* interpretation, that is, a fuzzy set modeling s_i is an entity or a *granula*. Moreover, we also use a *epistemic* interpretation, i.e., a fuzzy set is a model of non-precise data. Examples of observations with fuzzy modeling include: interval measurements [34], non-precise data [35], overlapping intervals described by

*Present address: IBM Research, Brazil. e-mail: jorged@br.ibm.com.
*Homepage: http://researcher.ibm.com/person/br-jorged.

[1]For example the inner product of two functions in RKHS's.

(subjective) words [36], a fuzzy version of precise data [37]. Fuzzy sets have a set of rich mathematical tools and they are a powerful modeling technique. They are widely used in, control systems, operations research, optimization, databases, etc[2]. Observations that are candidates to be modeled by fuzzy sets are: replicate measurements [38]; point-wise uncertainty [6]–[8]; subjective opinions [9]. meteorological, economics and bio-informatics data [7], [39]–[42].

Kernel embeddings. As we previously mentioned, kernels are mappings from an input space towards a geometrical space. This research focus on ML algorithms with either kernels embeddings of probability measures or kernels embeddings of fuzzy sets. That is the functional representation of probability measures and fuzzy sets in high dimensional spaces. Even though the kernel embedding of probability measures is not a new idea [12], [18], practical applications are still at its beginnings [4], [14], [43]–[49]. This thesis also presents the novel concept of kernel embeddings of fuzzy sets. This new idea will leverage several theoretical and practical applications on fuzzy sets.

A. Contributions

We list our principal theoretical and practical contributions within the areas of science benefited from them.

- *Fuzzy mathematics*: we proposed a new class of similarity measures between fuzzy sets via *kernels on fuzzy sets*. Those kernels make possible a *kernel embedding of fuzzy sets* into functional spaces. Hence, they enable a geometrical interpretation in high dimensional spaces, via the kernel trick, of several tasks involving fuzzy sets.
- *Fuzzy systems*: we proposed a new fuzzy system: *the non-singleton TSK fuzzy system*. We show that the dynamics of such fuzzy system can be viewed as a kernel. This view is useful to understand how functions are approximated by fuzzy systems in a RKHS context.
- *Kernel methods*: we formulate several classes of kernels on fuzzy sets: the intersection kernel, the cross product kernel, the convolution kernel, the distance-based kernel, the non singleton TSK kernel, all those kernels defined on fuzzy sets. We showed how to perform kernel engineering from those kernels. As result, we present the RBF, polynomial among others new kernels on fuzzy sets.
- *Fuzzy data analytics*: we applied kernels on fuzzy sets on supervised classification of athletic performance and dyslexic prediction [5]. We performed a hypotheses testing using a two-sample kernel test on clinical data using kernels on fuzzy sets [51].
- *Description models*: we propose three new non-parametric data description models for distributional data: the *support measure data description* models [52]–[54]. That is a kernel method, and it can be used as an one-class classifier. This method could be applied in classification, density estimation and clustering tasks. This method is based in kernel embeddings of probability measures.
- *Anomaly detection*: we applied description models for distributional data to the task of *group anomaly detection* on astronomical data [52], [54]. The task to be solved was given by finding out anomalous clusters of galaxies from a dataset of clusters of galaxies.
- *Machine learning and fuzzy machine learning*: we fill a gap between those two communities in the realm of kernel methods. There is no work analyzing fuzzy data from using kernels embeddings of fuzzy sets.

Our contributions were published in top conferences of the ML and Fuzzy sets areas: [5], [50]–[55].

B. Outline

A background in presented in Section II. The following sections are self-contained, each one describing our contributions, experiments and main results. We omitted proofs, however they are available in the thesis report or under request. Our main contributions are in the following sections: Section III formulates a data description model for distributional data. Section IV defines the kernels on fuzzy sets. Section V shows how fuzzy systems can be viewed from the kernel respective. Section VI shows how a distance between fuzzy sets can be used to generate new kernels on fuzzy sets. Section VII presents the conclusions.

II. Background

We briefly describe in this section kernels, fuzzy sets, probability measures, fuzzy data and distributional data.

A. Kernels

Kernels are real-valued functions defined on $\mathcal{X} \times \mathcal{X}$, where $\mathcal{X}$ is a non-empty set. They define similarity measures between objects or entities[3] [16], [18], [21]. Whereas a value $k(x, y)$ close to zero means that x and y are less similar. Higher values means more similarity between x and y. We say that k is positive definite if it satisfies:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(x_i, x_j) \geq 0, \tag{2}$$

for all $N \in \mathbb{N}$, $\{c_1, \ldots, c_N\} \subset \mathbb{R}$, $\{x_1, \ldots, x_N\} \subset \mathcal{X}$. If (2) has the inverse relation $\leq$, the kernel is called *negative definite*. If k is neither positive nor negative definite it is called *indefinite*. Positive definite kernels have the property

$$k(x, y) = \langle \phi_x, \phi_y \rangle_{\mathcal{H}}, \tag{3}$$

where $\mathcal{H}$ denotes a RKHS of functions $f : \mathcal{X} \to \mathbb{R}$, with kernel k, norm $\|.\|_{\mathcal{H}}$, and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Hence, ϕ_x, ϕ_y are functions in $\mathcal{H}$, and they are mappings of the form[4]:

$$\phi_x : \mathcal{X} \to \mathbb{R}, \quad y \mapsto \phi_x(y) = k(x, y). \tag{4}$$

We denote by $k(., s)$ the mapping $t \to k(t, s)$ with fixed s. Positive kernels are called *reproducing kernels* because they

[2] However, there is a lack of use of fuzzy modeling from the *core* ML community as it was noted in [37]. This research attempts to fill this gap in the kernel method area.

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

[4] Functions ϕ_x and ϕ_y are called as *representative functions* of x and y.

satisfy: 1) $\forall x \in \mathcal{X},\ \phi_x \in \mathcal{H}$ and 2) $\forall x \in \mathcal{X},\ \forall f \in \mathcal{H}$ $\langle f, \phi_x \rangle_{\mathcal{H}} = \langle f, k(.,x) \rangle_{\mathcal{H}} = f(x)$. The *reproducing property* of positive definite kernels is given by Condition 2). Equation (3) is derived from it.

Positive kernels provide a way to compute inner products in RKHS's for non-vectorial data using the kernel trick. Moreover, they permit to use geometrical tools and algorithms on non-vectorial data. For instance, to estimate empirical means, projections, angles, etc, on ϕ_y, ϕ_x.

Indefinite kernels either symmetric or non-symmetric had been used in ML problems with state-of-the-art results [56] [57]–[59]. Some work has been done to give a geometrical interpretation for indefinite kernels in pseudo-euclidean spaces [24], [56] and Krein spaces [23]. In such spaces, bilinear forms are not necessarily positive definite and norms do not define metrics. However, a geometric interpretation is given by linear spaces with symmetric bilinear forms[5]. For a deeper study of kernel methods we refer books [16], [18].

B. Fuzzy set Theory

Fuzzy sets are sets allowing membership degrees for their elements. A fuzzy set $X \in \Omega$ is a set characterized by a membership function $X : \Omega \to [0,1]$. Given $x \in \Omega$, the evaluation $X(x)$ is called as the *degree of membership of x to the fuzzy set X*. The *support* of a fuzzy set, denoted by $\mathrm{supp}(X)$, is the set $\{x \in \Omega \mid X(x) > 0\}$. Moreover, the set of all the fuzzy sets in Ω is denoted by $\mathcal{F}(\Omega)$. A intersection between fuzzy sets is ussually implement via a T-norm operator [61], [62]; which is a function $T : [0,1]^2 \to [0,1]$, such that, for all $x, y, z \in [0,1]$, satisfies: 1) commutativity: $T(x,y) = T(y,x)$; 2) associativity: $T(x, T(y,z)) = T(T(x,y), z)$; 3) monotonicity: $y \leq z \Rightarrow T(x,y) \leq T(x,z)$; and 4) limit condition $T(x,1) = x$. A multi-argument extension $T_n : [0,1]^n \to [0,1]$ is given by the recurrence:

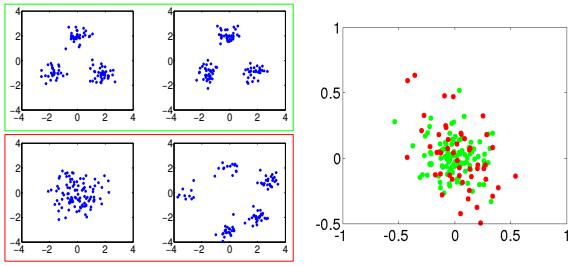
$$\begin{aligned} T_2(x_1, x_2) &= T(x_1, x_2), & n = 2, \\ T_n(x_1, x_2, \ldots, x_n) &= T(x_1, T_{n-1}(x_2, x_3, \ldots, x_n)), & n \geq 3. \end{aligned}$$

Throughout this text we use T to denote a T-norm and its multi-argument extension.

C. Probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a tuple with Ω being a non-empty set, $\mathcal{F}$ a σ-algebra and a *probability measure* $\mathbb{P} : \mathcal{F} \to [0,1]$. i.e. $\mathbb{P}$ satisfies the probability axioms. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*. We use notation $X \sim \mathbb{P}$ to say that the random variable X is distributed according to $\mathbb{P}$. Notation $\mathbb{E}_{X \sim \mathbb{P}}[f(X)]$ means the expectation of $f(X)$, where X is distributed according to $\mathbb{P}$. We use $\mathcal{B}(\mathbb{R}^D)$ to denote the Borel σ-algebra of $\mathbb{R}^D$.

[5]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 [60] for details.



(a) (b)

Figure 1. a) a) Red box: Two group anomalies. Green box: Two non-anomalous groups b) From several anomalous and non-anomalous groups similar to the ones from a) we compute the mean statistic per group. Red points are the means of anomalous groups. Blue points are the means of non-anomalous groups. Notice that it is hard it is to find out group anomalies using the mean statistic, because the overlapping.

D. Distributional data and Fuzzy data

We conclude this section defining distributional and fuzzy data [63]–[72]. By distributional data we mean that s_i is a sample according to a probability measure $\mathbb{P}_i$. In this sense (1) can be modeled by $\{\mathbb{P}_i\}_{i=1}^N$. By fuzzy data we mean either that s_i contains fuzzy elements. i.e., $s_i = (X_1, \ldots X_D)_i$, consequently (1) can be modeled by $\{(X_1, \ldots X_D)_i\}_{i=1}^N$ or s_i is a fuzzy set. i.e., (1) can be modeled by $\{X_i\}_{i=1}^N$.

III. Support Measure data Description

We now present in this section how to construct data description (DD) models for distributional data. DD models [28], [73]–[76] are useful in several ML tasks including one-class classification, clustering and description. We use our DD models in the task of group anomaly detection.

A. Contributions

Our main contributions are:

- a framework based on the concept of Minimum-Volume set (MV-set) to derive DD models for distributional data;
- a set of discriminative and non-parametric DD kernel methods for distributional data: the *support measure data description* (SMMD) models:
  - a SMDD model derived from minimum enclosing balls (MEB) of mean maps;
  - a SMDD model as the MEB of mean maps of norm one and stationary kernels;
  - a SMDD model as stochastic optimization problem.
- application of the models as one-class classifiers to solve the task of group anomaly detection on astronomical data;

B. Definitions

We present useful definitions regarding kernel embeddings [12], [12], [18], [45], [48], [77], [78] and Minimum-Volume (MV) sets [73]–[75].

Definition 1 (Mean map)**.** Let $\mathbb{P}$ be a probability measure and $X \sim \mathbb{P}$. The mean map in $\mathcal{H}$ is the function:

$$\begin{aligned} \mu_{\mathbb{P}} : \mathbb{R}^D &\rightarrow \mathbb{R} \\ t &\mapsto \mu_{\mathbb{P}}(t) = \mathbb{E}_{\mathbb{P}}[k(X,t)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x},t)d\mathbb{P}(\mathbf{x}), \end{aligned} \tag{5}$$

Definition 2. The embedding of probability measures $\mathbb{P} \in \mathcal{P}$ in $\mathcal{H}$ is given by the mapping

$$\begin{aligned} \mu : \mathcal{P} &\rightarrow \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X,.)] = \int_{\mathbf{x}\in\mathbb{R}^D} k(\mathbf{x},.)d\mathbb{P}(\mathbf{x}). \end{aligned}$$

Hence, $\mu_{\mathbb{P}}$ acts as the representative function in $\mathcal{H}$ for $\mathbb{P}$. The embedding is injective if the kernel k is characteristic [48], [79], [80]. Empirical estimators for $\mu_{\mathbb{P}}$, are bounded [12].

Theorem 1 (Kernel on probability measures)**.** *A real-valued kernel on $\mathcal{P} \times \mathcal{P}$, defined by*

$$\begin{aligned} \tilde{k}(\mathbb{P},\mathbb{Q}) =& \langle \mathbb{P},\mathbb{Q}\rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} \\ =& \int_{\mathbf{x}\in\mathbb{R}^D}\int_{\mathbf{x}'\in\mathbb{R}^D} k(\mathbf{x},\mathbf{x}')d\mathbb{P}(\mathbf{x})d\mathbb{Q}(\mathbf{x}') \end{aligned} \tag{6}$$

is positive definite [18].

Next, we present an extension of MV-sets to the case of probability measures

Definition 3 (MV-set for probability measures)**.** Let $(\mathcal{P},\mathcal{A},\mathcal{E})$ be a probability space, where $\mathcal{P}$ is the space of all probability measures $\mathbb{P}$ on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, $\mathcal{A}$ is some suitable σ-algebra of $\mathcal{P}$ and $\mathcal{E}$ is a probability measure on $(\mathcal{P},\mathcal{A})$. The MV-set is the set

$$G^*_\alpha = \underset{G\in\mathcal{A}}{\operatorname{argmin}}\{\rho(G) | \mathcal{E}(G) \geq \alpha\}, \tag{7}$$

where ρ is a reference measure on $\mathcal{A}$ and $\alpha \in [0,1]$. The MV-set G^*_α, describes a fraction α of the mass of $\mathcal{E}$.

We make the following assumptions: The i-th observation s_i contains i.i.d. [6] realizations of a random variable $X \sim \mathbb{P}_i$. The density and the form of each $\mathbb{P}_i$ is unknown. The sample $\{\mathbb{P}_i\}_{i=1}^N$ is i.i.d. according to $\mathcal{E}$ (Def. 3). The empirical measure: $\widehat{\mathbb{P}}_i = \frac{1}{L_i}\sum_{\ell=1}^{L_i}\delta_{\mathbf{x}_\ell}(s_i)$ approximates $\mathbb{P}_i$;

C. SMDD models

We introduce the *Support Measure Data Description Models* (SMDD) as a mean to describe distributional data. The formulation of the objective functions and the optimization problem of the SMDD models are given thanks to the definition of MV-set for probability measure. Our main idea is to use a minimum enclosing ball, in the RKHS, for the mean functions of probability measures and hence the distributional data. SMDD's solutions rely only on a subset of probability measures: the *support measures*.

[6]Independent and identically distributed.

1) Enclosing balls for volume sets : The definition of MV-sets is very general. Instead, we limit our attention to the class of sets $\mathcal{A}$ formed by sets of probability measures satisfying some specific criteria. A first empirical[7] approximation for G in (7) is given by:

$$\hat{G}_0(R,\mathbf{c}) = \{\mathbb{P}_i \in \mathcal{P} \mid \|X_i - \mathbf{c}\|^2 \leq R^2\}, \tag{8}$$

$R \in \mathbb{R}^+$ and center $\mathbf{c} \in \mathbb{R}^D$ are parameters of a hypersphere. A MV-set will be found optimizing over R and $\mathbf{c}$. However, (8) has two main drawbacks: it does not consider complex models, and some $\mathbb{P}_i$ will be in (8) if only if all possible realizations of $X_i \sim \mathbb{P}_i$ are inside the hypersphere $(R,\mathbf{c})$. Such limitations are overtaken considering the following three classes of sets:

- **Only mean maps inside the ball**

$$\hat{G}_1(R,c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c\|^2_{\mathcal{H}} \leq R^2\}, \tag{9}$$

- **mean maps with norm one and stationary kernels**,
- **a stochastic approach**

$$\hat{G}_3(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|k(X_i,.) - c\|^2_{\mathcal{H}} \leq R^2) \geq 1-\kappa_i\}. \tag{10}$$

The third class considers bounding values $\mathcal{K} = \{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0,1]$. Thus, $\mathbb{P}_i$ is in the volume-set G, if a subset of the realizations of the random variable $k(X,\cdot), X \sim \mathbb{P}_i$ is inside the hypersphere $(R,\mathbf{c})$, with probability less than $1-\kappa_i$.

D. First model SMDD

The MV-set $\hat{G}^*_\alpha$ for volume-sets G given by (9) can be computed solving the following optimization problem. Given $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ the SMDD model is:

Problem 1.

$$\begin{aligned} \min_{c\in\mathcal{H}, R\in\mathbb{R}^+, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & R^2 + \lambda\sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\mu_{\mathbb{P}_i} - c\|^2_{\mathcal{H}} \leq R^2 + \xi_i, i = 1,\ldots,N \\ & \xi_i \geq 0,\ i = 1,\ldots,N. \end{aligned}$$

Proposition 1 (Dual form)**.**

Problem 2.

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i,\mathbb{P}_i) - \sum_{i,j=1}^N \alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda,\ i = 1,\ldots,N \\ & \sum_{i=1}^N \alpha_i = 1 \end{aligned}$$

where $\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}\rangle_{\mathcal{H}}$ by (6)*, and $\boldsymbol{\alpha}$ is a Lagrange multiplier vector with non negative components α_i.*

Proposition 2 (Representer theorem)**.**

$$c(.) = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \leq \lambda\},$$

[7]Empirical in the sense of the sample $\{\mathbb{P}_i\}_{i=1}^N$.

where $\mathcal{I} = \{1, 2, \ldots, N\}$. *Furthermore, all* $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ *are inside the MV-set* $\hat{G}^*_\alpha$. *All* $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ *are the training errors. All* $\mathbb{P}_i,\ i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ *are the support measures.*

Theorem 2. *Let* η *be the Lagrange multiplier of the constraint* $\sum_{i=1}^N \alpha_i = 1$ *of Problem 2, then* $R^2 = -\eta + \|c\|^2_{\mathcal{H}}$.

A test probability measure $\mathbb{P}_t$ is in this SMDD model, if $\|\mu_{\mathbb{P}_t} - c\|^2_{\mathcal{H}}$ is a value less than R. The value $\|\mu_{\mathbb{P}_t} - c\|^2_{\mathcal{H}}$ is computed via

$$\tilde{k}(\mathbb{P}_t, \mathbb{P}_t) - 2\sum_i \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_t) + \sum_{i,j} \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j), \quad (11)$$

where indices i, j belongs to the support measure set.

E. Second SMDD Model

This SMDD model constraint means maps to have norm one. The main reason to do this is because stationary kernels [81]: $k_I(\mathbf{x}, \mathbf{x}') = f(\mathbf{x} - \mathbf{x}')$ have constant norm. However mean maps do not have constant norm [4]. This SMDD model performs the following normalization to means maps.

$$\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) = \frac{\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)}{\sqrt{\tilde{k}(\mathbb{P}_i, \mathbb{P}_i)\tilde{k}(\mathbb{P}_j, \mathbb{P}_j)}} = \frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}}, \quad (12)$$

which is actually a positive definite kernel. This normalization preserves the injectivity of the mapping $\mu : \mathcal{P} \to \mathcal{H}$. Therefore, the distributional information of the observations.

The MV-set $\hat{G}^*_\alpha$ for volume-sets G of the form given by (9), but with $\|\mu_{\mathbb{P}}\| = 1$ can be computed by solving Problem 2 but with kernel $\tilde{\tilde{k}}$. Furthermore, note that $\tilde{k}$ is given by (6) but with kernel k_I. As $\sum_{i=1}^N \alpha_i \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_i)$ is constant in Problem 2 when $\tilde{\tilde{k}}$ is used, the MV-set $\hat{G}^*_\alpha$ can be computed by the following optimization problem:

Problem 3.

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^N} \quad & -\sum_{i,j=1}^N \alpha_i \alpha_j \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda,\ i = 1, \ldots, N \\ & \sum_{i=1}^N \alpha_i = 1. \end{aligned}$$

This formulation is similar to the dual of the One-class Support Measures Machines [4], [75] but is not equivalent[8].

F. Third SMDD model

The MV-set $\hat{G}^*_\alpha$ for volume-sets G given by (10) can be computed by a chance-constrained optimization problem. Given $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$, and $\{\kappa_i\}_{i=1}^N,\ \kappa_i \in [0, 1]$, the SMDD model is:

Problem 4.

$$\begin{aligned} \min_{c \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \leq R^2 + \xi_i) \geq 1 - \kappa_i, \\ & \xi_i \geq 0, \\ & \text{for all} \ \ i = 1, \ldots, N. \end{aligned}$$

Chance constraints control the probability of constraint violation, allowing model flexibility. However, it is required to deal with every possible realization of $k(X, \cdot), X \sim \mathbb{P}_i$. Thus, it is necessary to turn probabilistic constraints into deterministic ones. Using Markov's inequality, we state that:

$$\mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}}[\|k(X_i, .) - c(.)\|^2_{\mathcal{H}}]}{R^2 + \xi_i}, \quad (13)$$

holds, for all $i = 1, 2, \ldots, N$.

The term $\mathbb{E}_{\mathbb{P}}[\|k(X_i, .) - c(.)\|^2_{\mathcal{H}}$ in (13) can be expressed in terms of the trace of a covariance operator in $\mathcal{H}$. A covariance operator in $\mathcal{H}$ with kernel k is the mapping $\Sigma^{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$, such that for all $f, g \in \mathcal{H}$ it satisfies:

$$\langle f, \Sigma^{\mathcal{H}} g \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[f(X)g(X)] - \mathbb{E}_{\mathbb{P}}[f(X)]\mathbb{E}_{\mathbb{P}}[g(X)],$$

because the reproducing property [9]. The covariance operator is then expressed in terms of tensorial products:

$$\Sigma^{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[k(X, .) \otimes k(X, .)] - \mathbb{E}_{\mathbb{P}}[k(X, .)] \otimes \mathbb{E}_{\mathbb{P}}[k(X, .)] \quad (14)$$

To compute the trace of covariance operator, we establish the following proposition and lemma.

Proposition 3.

$$tr(\Sigma^{\mathcal{H}}) = \mathbb{E}_{\mathbb{P}}[k(X, X)] - \tilde{k}(\mathbb{P}, \mathbb{P}). \quad (15)$$

That is, the trace of a possible infinite dimensional matrix can be computed in terms of kernel evaluations.

Lemma 1.

$$\mathbb{E}_{\mathbb{P}}[\|k(X, .) - c(.)\|^2_{\mathcal{H}}] = tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(.)\|^2_{\mathcal{H}}.$$

1) Deterministic Form: From Lemma (1), the deterministic form of the Problem 4 is the following optimization problem. Given the mean functions $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ and $\{\kappa_i\}_{i=1}^N,\ \kappa_i \in (0, 1]$, the SMDD model is:

Problem 5.

$$\begin{aligned} \min_{c \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} \leq (R^2 + \xi_i)\kappa_i - tr(\Sigma^{\mathcal{H}}_i), \\ & \xi_i \geq 0, \end{aligned}$$

for all $i = 1, \ldots, N$, where $tr(\Sigma^{\mathcal{H}}_i)$ is given by (15).

Proposition 4 (Dual form)**.** *The dual form of Prob. 5 is given by the following fractional programming problem*[10]*:*

[8]The thesis report contains a study of the equivalence between SMDD models and other discriminative models.

[9]$\Sigma^{\mathcal{H}}$ is a bounded operator on a separable infinite dimensional Hilbert space and can be represented by an infinite matrix [82].

[10]A reference for this kind of optimization problem is [83].

Problem 6.

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad \sum_{i=1}^{N}\alpha_i\langle\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_i}\rangle_{\mathcal{H}} - \frac{\sum_{i,j=1}^{N}\alpha_i\alpha_j\langle\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}\rangle_{\mathcal{H}}}{\sum_{i=1}^{N}\alpha_i} + \sum_{i=1}^{N}\alpha_i tr(\Sigma_i^{\mathcal{H}})$$

$$\text{subject to} \quad 0 \leq \alpha_i\kappa_i \leq \lambda,\ i=1,\ldots,N$$

$$\sum_{i=1}^{N}\alpha_i\kappa_i = 1,$$

where $\langle\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}\rangle_{\mathcal{H}}$ *is computed by* $\tilde{k}(\mathbb{P}_i,\mathbb{P}_j)$, $\boldsymbol{\alpha}$ *is a Lagrange multiplier vector with* α_i *non negative components; and* $tr(\Sigma_i^{\mathcal{H}})$ *is given by* (15).

Proposition 5 (Representer theorem)**.**

$$c(.) = \frac{\sum_i \alpha_i\mu_{\mathbb{P}_i}}{\sum_i\alpha_i},\ i\in\{i\in\mathcal{I}\mid 0<\alpha_i\kappa_i\leq\lambda\}, \tag{16}$$

where $\mathcal{I}=\{1,2,\ldots,N\}$. *Furthermore, all* $\mathbb{P}_i,\ i\in\{i\in\mathcal{I}\mid\alpha_i=0\}$ *are inside the MV-set* $\hat{G}^*_\alpha$. *All* $\mathbb{P}_i,\ i\in\{i\in\mathcal{I}\mid\alpha_i\kappa_i=\lambda\}$ *are the training errors. All* $\mathbb{P}_i,\ i\in\{i\in\mathcal{I}\mid 0<\alpha_i\kappa_i<\lambda\}$ *are the* support measures *and, from this, the radius is computed by*

$$R^2 = \frac{\|\mu_{\mathbb{P}_i}-c(.)\|^2 + tr(\Sigma_i^{\mathcal{H}})}{\kappa_i}, \tag{17}$$

for all $i\in\{i\in\mathcal{I}\mid 0<\alpha_i\kappa_i<\lambda\}$.

Alternatively, we have the following result to compute R.

Theorem 3. *Let* η *be the Lagrange multiplier of the constraint* $\sum_{i=1}^{N}\alpha_i\kappa_i=1$ *of the Lagrangian of Problem 6, then* $R^2=-\eta$.

In order to test if a probability measure $\mathbb{P}_t$ is described by this model, we have to compare R against $\|\mu_{\mathbb{P}_t}-c\|^2_{\mathcal{H}}+tr(\Sigma_t^{\mathcal{H}})$ which equals to (Prop. 5, Theorem 3, and Eq. (15))

$$\tilde{k}(\mathbb{P}_t,\mathbb{P}_t) - 2\sum_i\alpha_i\tilde{k}(\mathbb{P}_i,\mathbb{P}_t) + \sum_{i,j}\alpha_i\alpha_j\tilde{k}(\mathbb{P}_i,\mathbb{P}_j) + tr(\Sigma_t^{\mathcal{H}}) \tag{18}$$

G. Experiments on Supervised Group Anomaly Detection

We present some experiments on group anomaly detection[11].

1) Group anomaly detection: This task aims to find out anomalous groups of points from distributional data. Differently from usual anomaly detection in which anomalies are points far away of the center of the data, points of anomalous groups can be highly mixed with points of non-anomalous groups turning group anomaly detection a challenging problem. Group anomalies can be given by [1]: 1) *point-based* anomalies, defined as being an aggregation of anomalous points; 2) *distribution-based* anomalies, defined as being an anomalous aggregation of non-anomalous points. Our approach is to use SMDD models to find out group anomalies. Figure 1 shows why the information provided by each local distribution crucial to perform a right description of those datasets. For related works using generative and discriminative approaches we refer papers: [1], [2], [4], [84]–[86].

2) Experimental setting: We use empirical estimators to estimate the kernel between probability measures (6), and the the trace of the covariance operator (15). We used our three SMDD models and, for the sake of comparison, two state-of-the-art anomaly detection discriminative models. Table I shows the models and their respective notations. SVDD model was trained using the empirical group means. To see generative models vs OCSMM on the same problem we refer to [4] [12]. To get reliable statistics, we performed 200 runs. Performance metrics were area under the ROC curve (AUC), and accuracy (ACC). AUC values close to one indicate models find out group anomalies with few false positives and false negatives. A RBF kernel was always used within the kernel on probability measures. The kernel parameter was given by the median heuristic, that is one divided by the median of the Euclidean distance between all possible pairs of points in the dataset. The regularization parameter was set to one.

Model	Problem	Section/Ref.
M1	6	III-F
M2	2	III-D
M3	3	III-E
OCSMM	-	[4]
SVDD	-	[28]

Table I
MODELS USED IN THE EXPERIMENTS

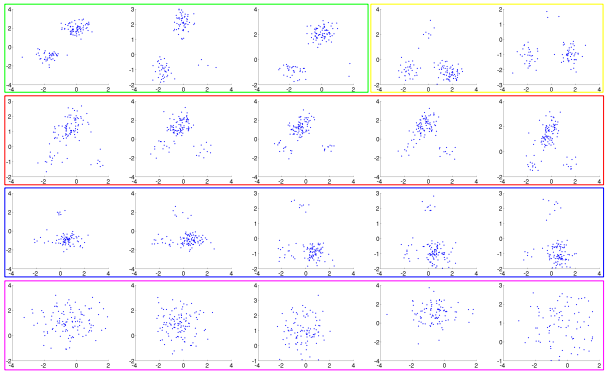
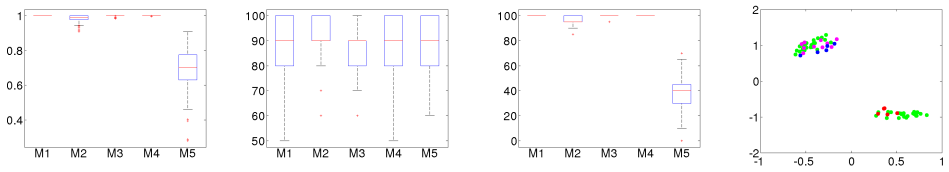


Figure 2. Group anomaly detection dataset. Green and yellow boxes contain non-anomalous groups of points. Red, blue, and magenta boxes contain anomalous groups of points.

3) Point-Based Group Anomaly Detection over a Gaussian Mixture Distribution dataset: We generated 50 non-anomalous groups and 30 groups for test. From the 30 groups in the test set, 20 groups correspond to anomalous groups. The number of points per group (anomalous or non-anomalous) follows a Poisson distribution with parameter $\beta = 10$. Points in non-anomalous groups were randomly sampled from two

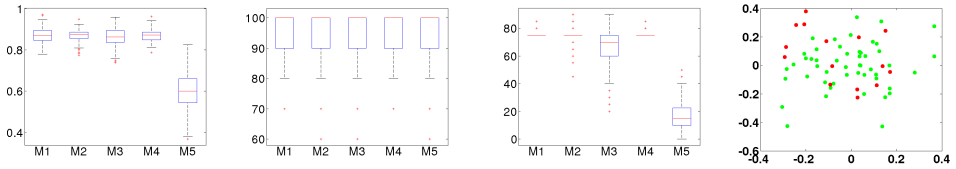
[11]More experiments on this data and artificial datasets were carried out and reported in the thesis manuscript.

[12]The Matlab code and datasets for experiments can be found at http://www.vision.ime.usp.br/~jorjasso/SMDD.html.



(a) AUC. (b) ACC Non A. (c) ACC Anom. (d) Group Means

Figure 3. Experimental results and a plot of the group means for the point-based group anomaly detection experiment.



(a) AUC. (b) ACC Non A. (c) ACC Anom. (d) Group Means

Figure 4. Experimental results and plot of the group means for the two distribution-based anomaly detection experiment.

different *Multimodal Gaussian Mixture Distribution* (GMD). The parameters for the first GMD were mixture weights: $(0.33, 0.64, 0.03)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$; and $0.2 * I_2$ as the sharing covariance matrix, where I_2 denotes the 2×2 identity matrix. The second GMD had the same parameters but mixture weights: $(0.33, 0.03, 0.64)$. The probability of chosen either the first or the second GMD was $\pi = (0.48, 0.52)$. The green box in Fig. 2 shows three non-anomalous groups for $\pi = 0.48$ and the yellow box shows two non-anomalous groups for $\pi = 0.52$.

We generated three different types of anomalous groups. The first type was given by 10 groups of points sampled from the normal distribution: $\mathcal{N}((-0.4, 1), I_2)$. Magenta box in Figure 2 show five of those groups. The second type was given by five groups of points sampled from GMD with the following parameters: weights: $(0.1, 0.08, 0.07, 0.75)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(0.6, -1)$; and a sharing covariance matrix given by $0.2 * I_2$. Blue box in Figure 2 shows five of those groups. The third type was given by five groups of points sampled from a GMD with parameters: weights: $(0.14, 0.1, 0.28, 0.48)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(-0.5, 1)$; and $0.2 * I_2$ as the sharing covariance matrix. Red box in Figure 2 shows five of those groups.

Figure 3 shows the metrics AUC, ACC for anomalous and non anomalous groups, for this experiment. To see the difficulty of the problem, we also plotted the group means of anomalous vs non-anomalous groups: Green points are the means of non-anomalous groups. Red, blue, and magenta points are the means of anomalous groups. Findings suggest that SMDD models can detect well such anomalies.

4) Distribution-Based Group Anomaly Detection over a Gaussian Mixture Distribution dataset: We generated 50 non-anomalous groups for the training set and 30 groups: 15 anomalous and 15 non-anomalous, for the test set. The number of points per group was the same as the last experiment.

Points from non-anomalous groups were sampled from a GMD with parameters: mixture weights: $p = \{1/3, 1/3, 1/3\}$; means: $(-1.7, 1), (1.7, -1), (0, 2)$ and sharing the same covariance matrix $0.2 * I_2$. The same GMD was used to generate group anomalies. However, the covariance matrix was given by estimating the covariance of the points rotated by 45 degrees.

Figure 4 shows the performance metrics for this experiment. In addition, we show a graph of the group means of non anomalous (green points) vs the anomalous groups (red points). Findings suggest that SMDD models perform well in this problem.

H. Group Anomaly Detection on Astronomical Data

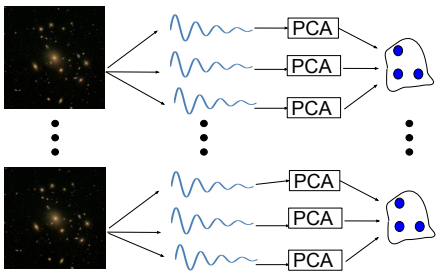


Figure 5. Feature extraction pipeline for the group anomaly experiment on astronomical data .

We show an application of SMDD models on astronomical data. The task is to find out anomalous groups of galaxies from a set containing several clusters of galaxies. Finding out those anomalies may help scientists to understand a common behavior among groups of galaxies in the universe. Also could leverage new discoveries by analyzing clusters with unusual behavior. The data was obtained from *The Sloan Digital Sky Survey*[13] *(SDSS)* project. Previous experiments with this dataset can be found in in [2]–[4]. This data contains massive spectroscopic surveys of the Milky Way galaxy and extra solar planetary systems. We used about 7×10^5 galaxies, each of them represented by a 4000-dimensional vector denoting

[13]http://www.sdss3.org/

spectral information. Following [3], each vector was down-sampled to a 500-dimensional vector. A neighborhood relation among galaxies was used to aggregate galaxies. The analysis returns 505 groups of galaxies from a total of 7530 galaxies. Thus, each group of galaxies contain about $10 - 15$ galaxies. Finally, we applied PCA to the spectral vectors, we used the first four PCA components, which preserved about 85% of the variance. Figure 5 shows the feature extraction pipeline

It is hard to have real examples of anomalous clusters of galaxies. For the sake of model comparison, we did two experiments. In both of them, we injected artificial group anomalies.

We set the first experiment by randomly using 455 non-anomalous groups as the training set. The test set was given by remaining 50 non-anomalous groups, plus 50 anomalous groups. We formed each group anomaly by randomly selecting n_i galaxies from all the dataset of galaxies. Parameter n_i was distributed according to a Poisson distribution with parameter $\beta = 15$. As galaxies were randomly chosen, the aggregation itself is anomalous.

The training set for the second experiment was given as before. We also used the remaining 50 non-anomalous groups plus 50 group anomalies. Each group anomaly was sampled from a GMD. We compute the mean, and the covariance of three random sets of galaxies. Those values was given as parameter of the GMD. The weights were $p = \{0.33, 0.33, 0.33\}$.

Figures 6d and 6h, show a graph of the group means of the four dimensional PCA vectors. Green points are the non-anomalous group means, and red points are the anomalous group means. From left to right, top to bottom, the 1st vs 2nd, 2nd vs 3rd , 3rd vs 4th and 4th vs 1st PCA dimension. Notice that group anomalies are hard to be detect by common methods, because there is an overlapping of group means.

Figure 6 shows performance metrics for the first experiment (top), and the second experiment (bottom). It is important to emphasize that OCSMM was compared against other group anomaly generative method and it obtained equivalent performance [4]. Therefore, we compare only SMDD models against OCSMM and SVDD models. The AUC metric shows that SMDD models performed well. Furthermore, a spherical normalization has a positive effect, increasing M3 AUC value close to one.

I. Summary and further research on this topic

We presented in this section the *Support Measure Data Description* (SMDD) models for distributional data. Some properties of those models are: 1) they are kernel methods; 2) the model is a function that only depends on a sub-set of probability measures: *the support measures*; 3) it is a non-parametric and discriminative model; 4) SMDD models are derived from MV-sets, when the class of functions are hyper-spheres in the RKHS. We successfully tested those models against the challenging task of group anomaly detection. We include a real application of detecting anomalous clusters of galaxies. Future work includes applications in novelty detection, clustering and classification, for distributional data, and extension to other learning methodologies.

IV. Kernels on fuzzy sets

Fuzzy sets similarity is an important topic of research due to its several theoretical and practical applications [87], [88]. We use a well-know concept from kernel methods, *the kernel*, to define a new class of similarity measures between fuzzy sets. In this sense, we present and define a new class of kernels: *kernels on fuzzy sets*. [5], [50], [51] They bring advantages and tools from kernels to the realm of fuzzy sets. Kernel on fuzzy sets implicitly define an embedding of fuzzy sets on functional spaces. Therefore, a geometrical interpretation of similarity measures for fuzzy sets. In the realm of ML and fuzzy data analytics, extending kernel methods to fuzzy data is straightforward, it only suffices to change the kernel definition in algorithms as support vector machines, kernel PCA, etc. Those advantages could leverage the development of many applications in several areas of research where the observational data is fuzzy data. Figure 7 shows a vectorial view of similarity measures for fuzzy sets in a RKHS.

Let $\mathcal{F}(\Omega) = \{X \mid X : \Omega \to [0,1]\}$ be the class of fuzzy sets on Ω. A kernel on fuzzy sets is a mapping of the form [5], [50], [51], [55]:

$$k : \mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \to \mathbb{R} \tag{19}$$

$$(X, Y) \mapsto k(X, Y), \tag{20}$$

Because reproducing property of kernels if k is positive definite, then there exist the following kernel embedding of fuzzy sets into a RKHS:

$$\phi : \mathcal{F}(\Omega) \to \mathcal{H}, \quad X \mapsto \phi_X(.) = k(., X). \tag{21}$$

Thus, similarity measures between fuzzy sets, via kernels on fuzzy sets, have the following geometrical view:

$$k(X, Y) = \langle \phi_X, \phi_Y \rangle_{\mathcal{H}}. \tag{22}$$

A. Contributions

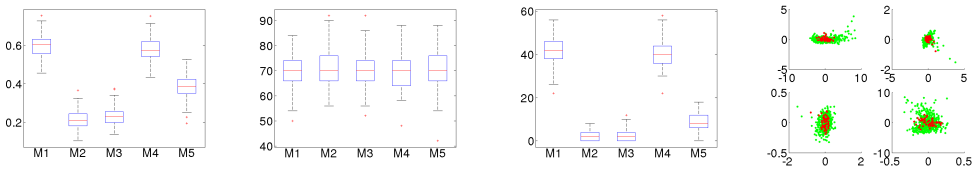
Our main contributions are :

- the development of the theory of behind kernels on fuzzy sets and kernel embeddings;
- the formulation of the intersection kernel on fuzzy sets;
- the formulation of the cross product kernel on fuzzy sets;
- the formulation of the convolution kernel on fuzzy sets;
- examples of applications on mathematics, ML and fuzzy data analytics

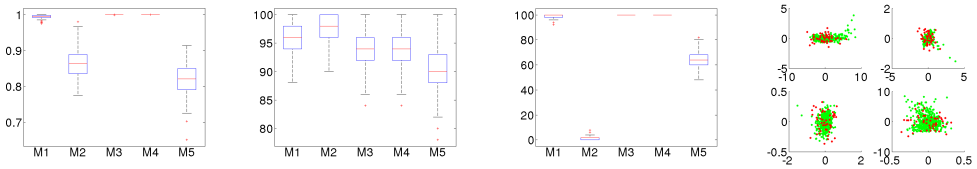
B. Related work

A related work using fuzzy sets and kernel methods, without implying kernels on fuzzy sets, are given by: clustering [89]–[91], classification [92], feature extraction [93], discriminant analysis [94]. Another related work linking fuzzy systems to kernels, without implying kernels on fuzzy sets are given by [95]–[103]. Also, there is a link between fuzzy equivalence relations and positive definite kernels [104]–[106]. To our best of knowledge, there is not a formulation of kernels on fuzzy sets.

Next, we present three class of kernels on fuzzy sets: the cross product, the intersection and the convolution kernel on fuzzy sets.



(a) AUC. (b) ACC Non A. (c) ACC Anom. (d) Group Means



(e) AUC. (f) ACC Non A. (g) ACC Anom. (h) Group Means

Figure 6. The results of the experiment for the group anomaly detection task over a SDSS III dataset.

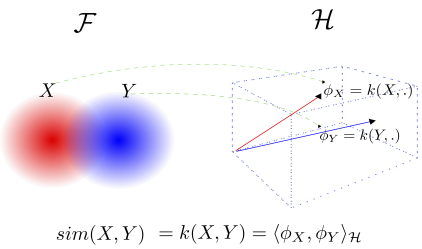


Figure 7. A geometrical view of a similarity measure between fuzzy sets using kernel on fuzzy sets.

C. The cross product kernel on fuzzy sets

1) The cross product kernel on sets: Let Ω be a nonempty set and Let $\mathcal{G}(\Omega)$ denote the set of all non-empty subsets of Ω. Assuming also $\mathcal{G}(\Omega)$ contains as elements all the nonempty *countable finite* subsets of Ω. The *cross product kernel* is the real-valued mapping $k : \mathcal{G}(\Omega) \times \mathcal{G}(\Omega) \to \mathbb{R}$ given by [107], [108]:

$$k_{set}(A, B) = \sum_{x \in A, y \in B} k(x, y), \tag{23}$$

where k is a real-valued kernel on $\Omega \times \Omega$. Kernel k_{set} defines a similarity measure for any two sets $A, B \in \mathcal{G}(\Omega)$ by $k(A, B) = \langle \phi_A, \phi_B \rangle_{\mathcal{H}}$, where ϕ_A and ϕ_B are the *representer functions* in a RKHS of the sets A, B, respectively.

Assuming that for all $X \in \mathcal{F}(\Omega)$ the set $\text{supp}(X)$ is a nonempty finite countable set, we present the cross product kernel on fuzzy sets in the following definition.

Definition 4 (The cross product kernel on fuzzy sets)**.** Given two real-valued kernels k_1, k_2 defined on $\Omega \times \Omega$ and $[0, 1] \times [0, 1]$ respectively. The cross product kernel on fuzzy sets is a function $k_\times : \mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \to \mathbb{R}$ given by:

$$k_\times(X, Y) = \sum_{\substack{x \in \text{supp}(X), \\ y \in \text{supp}(Y)}} k_1 \otimes k_2\big((x, X(x)), (y, Y(y))\big), \tag{24}$$

where $X(x)$ and $Y(y)$ are the membership degrees of $x, y \in \Omega$ to the fuzzy sets X, Y, and the *tensorial product*: $k_1 \otimes k_2 : (\Omega \times [0, 1]) \times (\Omega \times [0, 1]) \to \mathbb{R}$, is defined by:

$$k_1 \otimes k_2\big(x, X(x), y, Y(y)\big) = k_1(x, y)\ k_2(X(x), Y(y)). \tag{25}$$

Lemma 2. *If k_1 and k_2 are real-valued positive definite kernels, then $k_\times$ is a real-valued positive definite kernel.*

However, it is possible to use any kernels, positive or not, for k_1 and k_2 in $k_\times$. The geometrical representation for such cases is guaranteed [22], [23], [56], [60].

2) Examples:

Example 1. Table II shows several kernels $k_\times$ when $k_2\big(X(x), Y(y)\big) = X(x)Y(y)$, and k_1 is a positive definite kernel.

$k_1(x, y)$	$k_\times(X, Y)$
linear	$\sum_{\substack{x \in \text{supp}(X), \\ y \in \text{supp}(Y)}} xyX(x)Y(y)$
polynomial	$\sum_{\substack{x \in \text{supp}(X), \\ y \in \text{supp}(Y)}} (\alpha\langle x, y\rangle + \beta)^d X(x)Y(y)$
exponential	$\sum_{\substack{x \in \text{supp}(X), \\ y \in \text{supp}(Y)}} \exp(\sigma\langle x, y\rangle)X(x)Y(y)$
Gaussian	$\sum_{\substack{x \in \text{supp}(X), \\ y \in \text{supp}(Y)}} \exp(-\sigma\|x - y\|^2)X(x)Y(y)$

Table II
EXAMPLES OF KERNELS $k_\times$ FOR DIFFERENT FORMULATIONS FOR k_1 AND k_2 BEING THE LINEAR KERNEL.

The main difference between $k_\times(X, Y)$ and k_{set} is that the former treat each element from the set with equally importance.

Example 2. We explore the case when the elements to be analyzed are in the finite measure space $(\Omega, \mathcal{A}, \mu)$. Let k_1, k_2 be continuous functions with finite integral[14].

[14] k is a μ-integrable function, i.e., $k \in \mathcal{L}^1(\mu)$. (Definition 10.1, [109]).

The kernel

$$k_\times(X,Y) = \iint_{\substack{x\in\text{supp}(X),\\ y\in\text{supp}(Y)}} k_1\otimes k_2\big((x,X(x)),(y,Y(y))\big)d\mu(x)d\mu(y), \tag{26}$$

is a cross product kernel on fuzzy sets.

Example 3. If the data lives in a probability space $(\Omega,\mathcal{A},\mathbb{P})$, where $\mathbb{P}$ is a probability measure, the resulting cross product kernel on fuzzy sets will be

$$k_\times(X,Y) = \iint_{\substack{x\in\text{supp}(X),\\ y\in\text{supp}(Y)}} k_1\otimes k_2\big((x,X(x)),(y,Y(y))\big)d\mathbb{P}(x)d\mathbb{P}(y), \tag{27}$$

Moreover, if we have the set probability measures

$$\mathcal{P} = \{\mathbb{P} \mid \int_\Omega \sqrt{k_1(x,x)k_2\big(X(x),X(x)\big)}d\mathbb{P}(x) < \infty\}$$

The cross product kernel on fuzzy sets is given by:

$$k_\times(X,Y) = \iint_{\substack{x\in\text{supp}(X),\\ y\in\text{supp}(Y)}} k_1\otimes k_2\big((x,X(x)),(y,Y(y))\big)d\mathbb{P}(x)d\mathbb{Q}(y), \tag{28}$$

where elements from X and Y are i.i.d according to $\mathbb{P},\mathbb{Q}\in\mathcal{P}$.

Notice that in the last example we mix two types of uncertainty: randomness and fuzzyness (ontic or epistemic) Fuzzyness in the sense of membership degree of an element to a fuzzy set. Randomness in the sense of that the elements of such sets could follow some probabilistic law.

3) Generalization towards a product space: A generalization of $k_\times$ to deal with a D-tuple of fuzzy sets, i.e., $(X_1,\ldots,X_D)\in\mathcal{F}(\Omega_1)\times\cdots\times\mathcal{F}(\Omega_D)$ is implemented by the following kernel:

$$k_\times^\pi\big((X_1,\ldots,X_D),(Y_1,\ldots,Y_D)\big) = \prod_{d=1}^D k_\times^d(X_d,Y_d). \tag{29}$$

If all the kernels $k_\times^d$ are positive definite then $k_\times^\pi$ is positive definite by closure properties of kernels. Also, the following generalization is possible:

$$k_\times^\Sigma\big((X_1,\ldots,X_D),(Y_1,\ldots,Y_D)\big) = \sum_{d=1}^D \alpha_i k_\times^d(X_d,Y_d). \tag{30}$$

Kernel $k_\times^\Sigma$ is positive definite if only if $\alpha_i\in\mathbb{R}^+$ and all the $k_\times^D$ kernels are positive definite.

4) Properties: The following propositions show some properties of this kernel.

Proposition 6. *l $k_\times$ is a convolution kernel.*

Proposition 7. *$k_\times$ is a k_{set} and k_2 is a linear kernel over membership functions given by indicator functions.*

D. The fuzzy convolution kernel

We present a fuzzy version of the convolution kernel [110].

Definition 5 (The fuzzy convolution kernel)**.** Let e and $\vec{e}=(e_1,e_2,\ldots,e_L)$ elements of the sets E and $E_1\times\cdots\times E_L$, respectively. Given the fuzzy relation $R\subseteq(E_1\times E_2\times\cdots\times E_L)\times E$, with membership function:

$$\begin{aligned} R: E_1\times E_2\times\cdots\times E_L\times E &\to [0,1]\\ (\vec{e},e) &\mapsto R(\vec{e},e), \end{aligned} \tag{31}$$

and defining the decomposition $R^{-1}(e) = \{(\vec{e},R(\vec{e},e))|R(\vec{e},e)\}$. The fuzzy convolution kernel is a real-valued function on $E\times E$, satisfying $\forall e,e'\in E$:

$$k_\star(e,e') = \sum_{\substack{(\vec{e},R(\vec{e},e))\in R^{-1}(e),\\ (\vec{e}^{\,'},R(\vec{e}^{\,'},e'))\in R^{-1}(e')}} R(\vec{e},e)R(\vec{e}^{\,'},e')\prod_{l=1}^L k_l(e_l,e_l'), \tag{32}$$

where $k_l, 1\le l\le L$ are positive definite kernels on $E_l\times E_l$.

Lemma 3. *If kernels $k_1,k_2,\ldots,k_L$ are positive definite, then $k_\star$ is positive definite*

1) Examples:

Example 4. Consider the following fuzzy data:

$$\{x_i,\vec{x}_i,R(\vec{x}_i,x_i)\}_{i=1}^N, \tag{33}$$

where $x_i\in\Omega$ is an object, $\vec{x}_i = x_1^{(i)},\ldots x_L^{(i)}\in\Omega_1\times\cdots\times\Omega_L$ are a set of features and $R(\vec{x}_i,x_i)$ is a fuzzy relation modeling an membership degree of the features $\vec{x}_i$ to the object x_i. We use functions $f_l:\Omega_l\to\mathbb{R}^D$, $1\le l\le L$ e $D\in\mathbb{N}$, and the set of Gaussian kernels:

$$k_l(x_l^{(i)},x_l^{(j)}) = \exp(-\gamma_l\|f_l(x_l^{(i)})-f_l(x_l^{(j)})\|^2),\quad \gamma>0,$$

where $0\le i,j\le N$. Then, the following is a fuzzy convolution kernel on $\Omega\times\Omega$:

$$\begin{aligned} k_\star(x_i,x_j) = \quad & R(\vec{x}_i,x_i)R(\vec{x}_j,x_j)\\ & \exp(-\sum_l^L \gamma_l\|f_l(x_l^i)-f_l(x_l^j)\|^2). \end{aligned} \tag{34}$$

Example 5. Using $X,Y\in\mathcal{F}(\Omega)$ and the fuzzy relation:

$$\begin{aligned} R:\Omega\times\mathcal{F}(\Omega) &\to [0,1]\\ (\vec{x},X) &\mapsto R(\vec{x},X) = X(\vec{x}) \end{aligned}$$

The fuzzy convolution kernel is a kernel on $\mathcal{F}(\Omega)\times\mathcal{F}(\Omega)$:

$$\begin{aligned} k_\star(X,Y) &= \sum_{\substack{\vec{x},R(\vec{x},X)\in R^{-1}(X),\\ \vec{y},R(\vec{y},Y)\in R^{-1}(Y)}} R(\vec{x},X)R(\vec{y},Y)k(\vec{x},\vec{y})\\ &= \sum_{\substack{\vec{x},R(\vec{x},X)\in R^{-1}(X),\\ \vec{y},R(\vec{y},Y)\in R^{-1}(Y)}} X(\vec{x})Y(\vec{y})k(\vec{x},\vec{y}). \end{aligned} \tag{35}$$

This is the cross product kernel on fuzzy sets.

2) Generalization: Placing the kernel k_R on $[0,1]\times[0,1]$ in (32). i.e., $k_R\big(R(\vec{e},e),R(\vec{e}^{\,'},e')\big)$ instead of $R(\vec{e},e)R(\vec{e}^{\,'},e')$, give us a more general form for $k_\star$

E. The intersection kernel $k_\cap$ on fuzzy sets

This kernel is based on two key concepts:

- a *finite decomposition* of the support of a fuzzy set. In order to develop that concept we need the concept of semi-ring of sets;
- the concept of fuzzy contribution. We need here the definition of measure on semi-ring of sets

Next, we present some important definitions for the formulation of $k_\cap$.

1) Previous definitions:

Definition 6 (Semiring of sets)**.** A semi-ring of sets, $\mathcal{S} \subseteq \Omega$, is a subset of the power set $\mathcal{P}(\Omega)$, satisfying:

1 $\phi \in \mathcal{S}$, ϕ is the empty set,
2 $A, B \in \mathcal{S} \implies A \cap B \in \mathcal{S}$,
3 for all $A, A_1 \in \mathcal{S}$ such that $A_1 \subseteq A$, there is a sequence of pairwise disjoint sets: $A_2, A_3, \dots A_N \in \mathcal{S}$, satisfying:

$$A = \bigcup_{i=1}^{N} A_i.$$

Condition 3 is called as *finite decomposition of* A.

Definition 7 (Measure on semi-ring of sets)**.** Let $\mathcal{S}$ be a semi-ring of sets. Let $\rho : \mathcal{S} \to [0, \infty]$ be a function satisfying:

1 $\rho(\phi) = 0$;
2 $\rho(A) = \rho(\bigcup_{i=1}^{N} A_i) = \sum_{i=1}^{N} \rho(A_i)$. Where $A \in \mathcal{S}$ has finite decomposition: $\bigcup_{i=1}^{N} A_i$

by *Carathéodory's extension theorem* [111], ρ is a measure on $\sigma(\mathcal{S})$, where $\sigma(S)$ is the smallest σ-algebra containing S.

Using the both definitions, it is possible to define the **intersection kernel on sets.** [112]. This kernel is a function $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ defined by:

$$k(A, A') = \rho(A \cap A'). \tag{36}$$

This kernel is positive definite.

2) Measuring the support of fuzzy sets: We developed a way to measure the support of a fuzzy set. First, we define the set:

$$\mathcal{F}_\mathcal{S}(\Omega) = \{X \subset \Omega | supp(X) \in \mathcal{S}\},$$

Then, follows that, the support of any fuzzy set belonging to $\mathcal{F}_\mathcal{S}(\Omega)$, by Definition 6, has finite decomposition:

$$supp(X) = \bigcup_{A \in \mathcal{B} \subseteq \mathcal{A}} A. \tag{37}$$

Next, we introduce the following indicator function:

$$\begin{aligned} \mathbf{1}_B : \mathcal{S} &\to \{0, 1\} \\ A &\mapsto \mathbf{1}_B(A) = \begin{cases} 1, & A \subseteq B, \\ 0, & cc \end{cases}, \end{aligned} \tag{38}$$

where $B \subseteq \mathcal{S}$.

Definition 8 (Measure of the support of a fuzzy set)**.** Let $X \in \mathcal{F}_\mathcal{S}(\Omega)$ be a fuzzy set[15]. Let $\mathbf{1}_{supp(X)}$ be given by (38).

[15]Note that $supp(X)$ admits finite decomposition

We define a *measure of the support of a fuzzy set* being a measure on semi-ring of sets, satisfying:

$$\begin{aligned} \rho(supp(X)) &= \sum_{A \in \mathcal{B} \subseteq \mathcal{A}} \rho(A) && (39) \\ &= \sum_{A \in \mathcal{A}} \rho(A) \mathbf{1}_{supp(X)}(A). && (40) \end{aligned}$$

Example 6 (Measuring the support of an intersection of fuzzy sets)**.** Let X, Y be two fuzzy sets in $\in \mathcal{F}_\mathcal{S}(\Omega)$. A measure of the support of their intersection is given by:

$$\rho(supp(X \cap Y)) = \sum_{A \in \mathcal{A}} \rho(A) \mathbf{1}_{supp(X)}(A) \mathbf{1}_{supp(Y)}(A). \tag{41}$$

We present another key concept, the *fuzzy contribution*.

Definition 9 (Fuzzy contribution)**.** Let X be a fuzzy set in $\in \mathcal{F}_\mathcal{S}(\Omega)$. The fuzzy contribution of the set $A \in \mathcal{S}$ is the value:

$$X(A) = \sum_{x \in A} X(x). \tag{42}$$

3) The intersection kernel on fuzzy sets: Based on the above definitions, we state the intersection kernel on fuzzy sets in the following definition.

Definition 10 (The kernel $k_\cap$)**.** Let X, Y be two fuzzy sets in $\mathcal{F}_\mathcal{S}(\Omega)$. We define the kernel $k_\cap$ as being the function: $k_\cap : \mathcal{F}_\mathcal{S}(\Omega) \times \mathcal{F}_\mathcal{S}(\Omega) \to \mathbb{R}$, satisfying:

$$k_\cap(X, Y) = \sum_{A \in \mathcal{A}} \big(X \cap Y\big)(A) \rho(A) \mathbf{1}_{supp(X)}(A) \mathbf{1}_{supp(Y)}(A), \tag{43}$$

Where $\big(X \cap Y\big)(A)$ is the fuzzy contribution of $A \in \mathcal{A}$.

For the sake of easy notation, we introduce: $\mathcal{C}_{X,Y} = \{A \in \mathcal{A} | \mathbf{1}_{supp(X)}(A) \mathbf{1}_{supp(Y)}(A) = 1\}$. Then, follows that (43) can be rewritten as:

$$k_\cap(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} X \cap Y(A) \rho(A). \tag{44}$$

4) Examples:

Example 7. Using a T-norm $k_\cap$ can be rewritten as:

$$\begin{aligned} k_\cap(X, Y) &= \sum_{A \in \mathcal{C}_{X,Y}} X \cap Y(A) \rho(A) && (45) \\ &\quad \text{by fuzzy contribution (Definition 9)} \\ &= \sum_{A \in \mathcal{C}_{X,Y}} \left(\sum_{x \in A} X \cap Y(x) \right) \rho(A) \\ &\quad \text{using a T-norm} \\ &= \sum_{A \in \mathcal{C}_{X,Y}} \left(\sum_{x \in A} T(X(x), Y(x)) \right) \rho(A) && (46) \end{aligned}$$

Example 8. Table III shows several kernels $k_\cap(X, Y)$ derived from common T-norms.

Example 9. If the σ-algebra for ρ is a Borel algebra, then $k_\cap$ can be rewritten as:

$$k_\cap(X, Y) = \sum_{A \in \mathcal{C}_{\mathcal{X},\mathcal{Y}}} \int_{x \in A} X \cap Y(x) d\rho(x) \equiv \int_{x \in \Omega} X \cap Y(x) d\rho(x), \tag{47}$$

Kernel $k_\cap$	T-norm
$k_{\cap\_\min}(X,Y) = \sum_{A\in\mathcal{C}_{\mathcal{X},\mathcal{Y}}} \sum_{x\in A} \min(X(x),Y(x))\rho(A)$	minimum
$k_{\cap\_\text{pro}}(X,Y) = \sum_{A\in\mathcal{C}_{\mathcal{X},\mathcal{Y}}} \sum_{x\in A} X(x)Y(x)\rho(A)$	product
$k_{\cap\_\text{Łuk}}(X,Y) = \sum_{A\in\mathcal{C}_{\mathcal{X},\mathcal{Y}}} \sum_{x\in A} \max(X(x)+Y(x)-1,0)\rho(A)$	Łukasiewicz
$k_{\cap\_\text{Dra}}(X,Y) = \sum_{A\in\mathcal{C}_{\mathcal{X},\mathcal{Y}}} \sum_{x\in A} Z(X(x),Y(x))\rho(A)$	Drastic

Table III
DIFFERENT FORMULATIONS FOR $k_\cap$ INDUCED BY DIFFERENT T-NORMS

Example 10. From the last example, if the measure is the probability measure $\mathbb{P}$, then $k_\cap$ can be rewritten as:

$$k_\cap(X,Y) = \int_{\mathbb{R}^D} T(X(x),Y(x))d\mathbb{P}(x). \tag{48}$$

Specific T-norms, for instance, the minimum and product T-norms produce the following kernels:

$$k_{\cap\_\min}(X,Y) = \int_{\mathbb{R}^D} \min(X(x),Y(x))d\mathbb{P}(x) \tag{49}$$

$$k_{\cap\_\text{pro}}(X,Y) = \int_{\mathbb{R}^D} X(x)Y(x)d\mathbb{P}(x) \tag{50}$$

5) Positivity of the kernels: We have proved the following propositions showing that kernels $k_{\cap\_\min}$ and $k_{\cap\_\text{pro}}$ are reproducing kernels.

Proposition 8. *Kernel $k_{\cap\_\min}(X,Y)$ is positive definite.*

Proposition 9. *Kernel $k_{\cap\_pro}(X,Y)$ is positive definite.*

Theorem 4 (A positive definite intersection kernel on fuzzy sets)**.** *If the T-norm T is a positive definite function, then $k_\cap$ is positive definite.*

F. Mathematical applications

1) Kernel embedding of fuzzy sets into a RKHS: Positive definite kernels on fuzzy sets perform, implicitly, the following embedding of fuzzy sets into a RKHS:

$$\phi : \mathcal{F}(\Omega) \to \mathcal{H}, \quad X \mapsto \phi_X(.) = k(.,X).$$

That mapping is a nonlinear one. An important result of this embedding is that it is possible to have a vectorial representation of fuzzy sets in a RKHS.

2) Distances on fuzzy sets: If the mapping is injective, then the following expression is a metric in a RKHS:

$$D(X,Y) \stackrel{\text{def}}{=} \|k(.,X) - k(.,Y)\|_{\mathcal{H}}. \tag{51}$$

That metric induce a topology on the set of fuzzy sets $\mathcal{F}(\Omega)$. Moreover, relaxing the condition it is possible to induce pseudo-metrics and semi-metrics in a RKHS. Both of them inducing a topology in the set $\mathcal{F}(\Omega)$.

3) A new geometrical view of fuzzy sets: Kernels on fuzzy sets leverages a new geometrical view of fuzzy sets. For example, for two fuzzy sets X, Y, it is possible to estimate:

$$\langle \phi_X, \phi_Y \rangle_{\mathcal{H}} = \|\phi_X\|_{\mathcal{H}} \|\phi_Y\|_{\mathcal{H}} \cos\theta = k(X,Y)$$

4) A new similarity notion for fuzzy sets: Kernel values will be within:

$$0 \leq k(X,Y) \leq \|\phi_X\|_{\mathcal{H}} \|\phi_Y\|_{\mathcal{H}}.$$

In the extreme case, we have either, $k(X,Y) = \|\phi_X\|_{\mathcal{H}} \|\phi_Y\|_{\mathcal{H}}$ if the angle between ϕ_X and ϕ_Y is zero; or $k(X,Y) = 0$ if ϕ_X and ϕ_Y are orthogonal.

5) Open questions: An important open question is about the characterization of injective mappings ϕ. Specifically, what kind of kernels on fuzzy sets make ϕ to be injective.

G. Kernel engineering

Let k_1, k_2 be two positive definite kernels on fuzzy sets. Value $\alpha \in \mathbb{R}^+$, function $f : \mathcal{F}(\Omega) \to \mathbb{R}$, a positive definite kernel k_3 on $\mathbb{R}^D \times \mathbb{R}^D$, a $N \times N$ positive definite symmetric matrix Σ and $\phi : \mathcal{F}(\Omega) \to \mathbb{R}^D$. By closure properties of positive kernels [17], the following are positive definite kernels on fuzzy sets 1) $k = k_1 + \alpha$; 2) $k = k_1 + k_2$; 3) $k = \alpha k_1$; 4) $k(x,x') = k_1(x,x')k_2(x,x')$; 5) $k(x,x') = f(x)f(x')$; 6) $k(x,x') = k_3(\phi(x),\phi(x'))$. Based on those primitives, we have the following positive definite kernels on fuzzy sets:[16]

- The convex combination kernel on fuzzy sets:

$$k_{mkl}(X,X') = \sum_{m=1}^{M} \alpha_m k_m(X,X'), \text{ com } \alpha_m \in \mathbb{R}^+ \tag{52}$$

- The polynomial kernel on fuzzy sets:

$$k_{pol}(X,Y) = (k(X,Y)+\alpha)^\beta, \ \alpha \geq 0, \beta \in \mathbb{N} \tag{53}$$

- The RBF kernel on fuzzy sets:

$$k_{gauss}(X,Y) = \exp(-\gamma D(X,Y)), \ \gamma > 0 \tag{54}$$

- The rational quadratic kernel on fuzzy sets

$$k_{ratio}(X,Y) = (1 + \frac{D(X,Y)}{\alpha\beta^2})^{-\alpha}, \ \alpha,\beta > 0 \tag{55}$$

Also, it is possible to have the following *Conditionally positive definite* kernels

- Multiquadric kernel on fuzzy sets:

$$k_{multi}(X,Y) = -\sqrt{D(X,Y)+\alpha^2}, \ \alpha \in \mathbb{R} \tag{56}$$

- Inverse multiquadric kernel on fuzzy sets.

$$k_{invmult}(X,Y) = (\sqrt{D(X,Y)+\alpha^2})^{-1}, \ \alpha \in \mathbb{R} \tag{57}$$

[16]All this examples assume that D is a metric on fuzzy sets.

H. Fuzzy data analytics and ML applications

Kernels o fuzzy sets will be used in modular form in kernel methods to solve several ML problems and fuzzy data analytics. Methods as support vector machines, Fisher discriminant analysis [113], kernel logistic regression [114], among others can be extended for the analysis of fuzzy data. Approaches as Multiple Kernel Learning [30] can be used for the case of heterogeneous data involving fuzzy variables. For instance, a *regression task* can be extended to include fuzzy set modeling for imprecise inputs. i.e., the regression function $\mathbb{E}[Y|X = X_i]$ takes into account not only the uncertainty of the dependent variable but also of the independent variable. Some methods as the support vector regression [115] and the Gaussian proccess [29] can be easily adapted to this case. Data description models the support vector data description method [116] and the one-class support vector machine could be easily extended to the case of fuzzy data. The only thing to be modified will be a kernel matrix induced by kernels on fuzzy sets. Those methods could leverage applications as anomaly detection on fuzzy data, density estimation for imprecise data, etc.

I. Summary and further research on this topic

We defined kernel on fuzzy sets. We present the cross product kernel, the convolution kernel and the intersection kernel on fuzzy sets. All of them are similarity measures for fuzzy sets. We presented the theory behind them and several examples. Also, we list several areas that could benefit from those kernels.

V. Takagi Sugeno Kernel

We present in this section a new class of kernels on fuzzy set derived from an extension of a Takagi Sugeno Kang (TSK) fuzzy systems [117] for the case of non-singleton inputs. We call those kernels as *non-singleton TSK kernel on fuzzy sets*

Interpretation of fuzzy systems as kernels is important because this give a new view of the representation of the space of functions whereby fuzzy systems approximate hypothesis. For instance, if the fuzzy system induce a positive definite kernel, the space of functions that the system approximates is a RKHS. Besides linguistic interpretation, positive definite kernels give an additionally property to fuzzy systems: a geometrical view in a RKHS of the interaction of fuzzy rules and inputs to the system.

A. Related work

We classify the related work as follows: 1) *fuzzy theory and kernels.* Some work using jointly positive kernels and fuzzy sets are: [89]–[94]. 2) *Fuzzy sets inducing kernels.* Fuzzy systems with some T-norm operators induce positive define kernels. [95], [96]. Some applications using this fact to train fuzzy learning systems using support vector machines [97]–[102]. 3) *Kernels and fuzzy basis functions.* A related work linking fuzzy basis functions and kernels are: [118]. 4) *Kernels and fuzzy equivalence relations.* Literature shows that some positive kernels can be interpreted as fuzzy equivalence relations [104], [105]. The importance of this is that some prior knowledge can be introducing in the algorithm [106]. None of the related work imply kernels on fuzzy sets.

B. Contributions

We have the following contributions:

- we define the *non-singleton TSK fuzzy system*;
- we show that this new fuzzy system induce a class of kernels: the *non-singleton TSK kernels*;
- we give several examples of implementation of TSK kernels;
- we show that non-singleton TSK kernels are fuzzy equivalence relations;
- we give two applications on a supervised classification on imprecise datasets: an athletics performance and a dyslexia prediction.

C. TSK fuzzy system

TSK fuzzy systems [117], are fuzzy systems that describe the outputs of the rules by functions. Given the index sets $\mathcal{D} = \{1, \ldots, d, \ldots, D\}$ e $\mathcal{L} = \{1, \ldots, \ell, \ldots, L\}$ a TKS rule system is a set with L elements or *rules* of the form:

$$\underbrace{\textbf{If } X_1^\ell \text{ and} \ldots \text{and } \textbf{If } X_d^\ell \text{ and} \ldots \text{and } \textbf{If } X_D^\ell}_{\text{antecedent}} \qquad \underbrace{\textbf{Then } g^\ell(\mathbf{x})}_{\text{consequent}}, \tag{58}$$

where X_d^ℓ is a fuzzy set in $\mathcal{F}(\Omega_d)$, Ω_d is a domain, and g^ℓ is a real-valued function on $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_D$.

Definition 11. The antecedent part of the rule can be modeled by the following fuzzy relation:

$$\begin{aligned} A^\ell : \prod_{d=1}^{D} \Omega_d &\to [0,1] \\ \mathbf{x} &\mapsto A^\ell(\mathbf{x}) = T(X_1^\ell(x_1), X_2^\ell(x_2), \ldots, X_D^\ell(x_D)). \end{aligned} \tag{59}$$

$\forall \ell \in \mathcal{L}$, where $\mathbf{x}$ is the tuple $(x_1, x_2, \ldots, x_D)$.

This allows to represent the rules by the following mappings

Definition 12. A rule $\ell \in \mathcal{L}$ of a TSK fuzzy system (58) is modeled by the mapping:

$$\begin{aligned} \prod_{d=1}^{D} \Omega_d &\to \mathbb{R} \\ \mathbf{x} &\mapsto A^\ell(\mathbf{x}) g^\ell(\mathbf{x}). \end{aligned} \tag{60}$$

A TSK fuzzy system define real-valued functions on $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_D$ by:

$$f_s(\mathbf{x}) = \frac{\sum_{\ell=1}^{L} A^\ell(\mathbf{x}) g^\ell(\mathbf{x})}{\sum_{\ell=1}^{L} A^\ell(\mathbf{x})}. \tag{61}$$

D. Non-singleton TSK fuzzy system

Instead of modeling the inputs to the system by $\mathbf{x}$, we will use a tuple of fuzzy sets $(X_1, X_2, \ldots, X_D)$. In the context of Mamdani fuzzy systems, such systems are called non-singleton fuzzy systems [119].

Definition 13 (TSK-NS fuzzy system)**.** Let A^ℓ and g^ℓ be mappings as defined before. A TSK-NS is a fuzzy system with L rules given by (58), defining a real-valued mapping on $\mathcal{F}(\Omega_1 \times \Omega_2 \times \cdots \times \Omega_D)$, given by:

$$f_{ns}(X; \mathbf{x}'') = \frac{\sum_{\ell=1}^{L} \sup_{\mathbf{x}\in\Omega_1\times\cdots\times\Omega_D} \big(T(X(\mathbf{x}), A^\ell(\mathbf{x}))\big)\, g^\ell(\mathbf{x}'')}{\sum_{\ell=1}^{L} \sup_{\mathbf{x}\in\Omega_1\times\cdots\times\Omega_D} T(X(\mathbf{x}), A^\ell(\mathbf{x}))}, \tag{62}$$

where $\mathbf{x}'' \in \prod_{d=1}^{D} \Omega_d$ is a parameter and $\mathcal{F}(\Omega_1 \times \cdots \times \Omega_D)$ is the class of fuzzy sets defined on $\Omega_1 \times \cdots \times \Omega_D$.

E. kernels and fuzzy systems

We review related work with fuzzy systems and rules.

1) PDFC Kernel [95]: This kernel is a real-valued kernel on $\mathbb{R}^D \times \mathbb{R}^D$. It can be induced by Mamdani and TSK fuzzy systems.

Lemma 4 (PDFC kernel)**.** *If for all $d \in \mathcal{D}$ and $\ell \in \mathcal{L}$, any fuzzy set X_d^ℓ, of the antecedent part of the rule* (58) *were given by translating the following positive definite functions: $u_d : \mathbb{R} \to [0, 1],\ d \in \mathcal{D}$, given by:*

$$u_d(x) = \begin{cases} u_d(-x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0, \end{cases}, x \in \mathbb{R}, \tag{63}$$

that is, $X_d^\ell(x_d) = u_d(x_d - x_d^\ell)$, where x_d is a particular input and x_d^ℓ its corresponding dislocation parameter. And if it is used a product T-norm in (59)*, then the kernel $k_{pdfc} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, given by:*

$$\begin{aligned} k_{pdfc}(\mathbf{x}, \mathbf{x}^\ell) &= \prod_{d=1}^{D} u_d(x_d - x_d^\ell) \\ &= \prod_{d=1}^{D} X_d(x_d) \\ &= A^l(\mathbf{x}), \end{aligned} \tag{64}$$

is positive definite.

2) Kernel Takagi Sugeno Kang (TSK) [99], [102]: If the following conditions are satisfied: 1) a Gaussian functions for antecedents fuzzy sets; 2) a product function as T-norm operator; 3) a consequent of the form $g^\ell(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}^\ell \rangle$, where $\mathbf{x}^\ell$ is a parameter of g^ℓ; then, the following kernel: $k_{tsk} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ given by:

$$\begin{aligned} k_{tsk}(\mathbf{x}, \mathbf{x}^\ell) &= \prod_{d=1}^{D} \exp(-\gamma(x_d - x_d^\ell)^2) \langle \mathbf{x}, \mathbf{x}^\ell \rangle \\ &= \exp(-\gamma ||\mathbf{x} - \mathbf{x}^\ell||^2) \langle \mathbf{x}, \mathbf{x}^\ell \rangle \\ &= A^\ell(\mathbf{x}) g^\ell(\mathbf{x}) \end{aligned}$$

where $\gamma \in \mathbb{R}^+$, is positive definite.

We have the following lemma generalizing such a result.

Lemma 5 (Kernels TSK)**.** *Let $\mathbf{x} \in \mathbb{R}^D$ be an input to a TSK fuzzy system. Let $\mathbf{x}^\ell \in \mathbb{R}^D$ be parameters of fuzzy sets of the antecedent part of the rules.. If A^ℓ and g^ℓ were given by: $A^\ell(\mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}^\ell),\ g^\ell(\mathbf{x}) = k_2(\mathbf{x}, \mathbf{x}^\ell)$, where k_1 and k_2 are positive definite kernels. Then, the function defined by Definition 12, is a positive definite kernel.*

F. Non-singleton TSK kernel on fuzzy sets

We present a new class of kernels on fuzzy sets induced by a non-singleton TSK fuzzy system. For the sake of notation, we use notation Ω for the Cartesian product $\Omega_1 \times \ldots \Omega_D$. We denote by Y the fuzzy set A^ℓ. In this sense, expression $T(X(\mathbf{x}), Y(\mathbf{x}))$ always means $T(X(\mathbf{x}), A^\ell(\mathbf{x}))$.

From definition of non-singleton TSK fuzzy system (Definição 13), specifically rule's antecedents, we have the following kernel on fuzzy sets.

Definition 14 (Non-singleton TSK Kernel)**.** A non-singleton TSK kernel on fuzzy sets is the function $\mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \to [0, 1]$ defined by:

$$k_{ntsk}(X, Y) = \sup_{\mathbf{x}\in\Omega} \left(T(X(\mathbf{x}), Y(\mathbf{x}))\right). \tag{65}$$

As T-norms implement intersection of fuzzy sets. This kernel is a kernel based on the intersection of fuzzy sets. i.e., $k_{ntsk}(X, Y) = \sup_{\mathbf{x}\in\Omega} \big(X \cap Y\big)(\mathbf{x})$.

Theorem 5. *If $\mathcal{F}(\Omega)$ contain only normal fuzzy sets, then k_{ntsk} is positive definite.*

Lemma 6. *Kernel k_{ntsk} can be written as:*

$$\begin{aligned} k_{ntsk}(X, Y) = T\Bigg(&\sup_{x_1\in\Omega_1} T\left(X_1(x_1), Y_1(x_1)\right), \ldots, \\ &\sup_{x_D\in\Omega_D} T\left(X_D(x_D), Y_D(x_D)\right)\Bigg) \end{aligned} \tag{66}$$

Moreover, using the rule's consequents, we have the following real-valued kernel on fuzzy sets.

Definition 15 (Non-singleton TSK Kernel with consequent)**.** This kernel is the function: $k_{ntsk} \otimes g^\ell : \big(\mathcal{F}(\Omega) \times \Omega\big) \times \big(\mathcal{F}(\Omega) \times \Omega\big) \to [0, 1]$, given by:

$$k_{ntsk} \otimes g^\ell\big((X, \mathbf{x}), (Y, \boldsymbol{\theta})\big) = k_{ntsk}(X, Y) g^\ell(\mathbf{x}; \boldsymbol{\theta}) \tag{67}$$

This kernel is induced by a TSK-NS fuzzy system, where: $g^\ell(., \boldsymbol{\theta}) : \Omega \to \mathbb{R}$ is the consequent function of rule ℓ with parameters $\boldsymbol{\theta} \in \Omega$.

Lemma 7. *$k_{ntsk} \otimes g^\ell$ is a positive definite kernel if k_{ntsk} and g^ℓ are positive definite kernels.*

G. Non-singleton TSK Gaussian kernel

Setting Gaussian membership functions for all the fuzzy sets: i.e.,: $\forall d \in \mathcal{D},\ X_d \in \mathcal{F}(\mathbb{R})$ are defined by:

$$X_d(x_d) = \exp\left(-\frac{1}{2}\frac{(x_d - m_d)^2}{\sigma_d^2}\right),$$

where, $x_d \in \mathbb{R}$ and $m_d \in \mathbb{R}$ e $\sigma_d \in \mathbb{R}^+$ are parameters. By T-norm, $X \in \mathcal{F}(\Omega)$ is given by

$$X(\mathbf{x}) = \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(x_d - m_d)^2}{\sigma_d^2}\right).$$

Thus, a non-singleton TSK Gaussian kernel for the tuples of fuzzy sets $X = (X_1, \ldots, X_D)$ and $Y = (Y_1, \ldots, Y_D)$, where each fuzzy set X_d (or Y_d) have parameters (m_d, σ_d) (or $(, m'_d, \sigma')_d)$), is given by:[17]

$$\begin{aligned} & k_{ntsk}(X, Y) \\ = & \sup_{\mathbf{x}\in\mathbb{R}^D} (T(X(\mathbf{x}), Y(\mathbf{x}))) \\ = & \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(x_d^{sup} - m_d)^2}{\sigma_d^2} - \frac{1}{2}\frac{(x_d^{sup} - m'_d)^2}{(\sigma')_d^2}\right). \end{aligned}$$

Values x_d^{sup} can be computed by taking the derivative of the above expression with respect to x_d^{sup} and setting this expression to zero. This procedure gives the value:

$$x_d^{sup} = \frac{m_d(\sigma'_d)^2 + m'_d\sigma_d^2}{\sigma_d^2 + (\sigma'_d)^2}. \tag{68}$$

Summarizing all the above expression we have the following definition.

Definition 16 (non-singleton TSK Gaussian kernel)**.** The non-singleton TSK Gaussian kernel on fuzzy sets is given by

$$k_{ntsk}(X, Y) = \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(m_d - m'_d)^2}{\sigma_d^2 + (\sigma'_d)^2}\right), \tag{69}$$

Corollary 1. *the non-singleton TSK Gaussian kernel is a positive definite kernel.*

This kernel is positive definite thanks to Theorem 5. Additionally, it is possible to parametrize the kernel with a bandwidth parameter.

Definition 17. The non-singleton TSK Gaussian kernel with parameter $\gamma \in \mathbb{R}^+$ is a kernel on fuzzy sets given by

$$k_{ntsk}^{\gamma}(X, Y) = \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(m_d - m'_d)^2}{\sigma_d^2 + (\sigma'_d)^2 + \gamma}\right), \tag{70}$$

Notice that this kernel is positive definite, because changing σ''_d by $\sigma'_d + \gamma$ in (70), we get (69).

Moreover, it is possible to induce several non-singleton TSK Gaussian kernels from Definition 15 by defining an appropriate consequent function. The following expressions are non-singleton TSK Gaussian kernels with consequents[18] i.e., $k_{ntsk} \otimes g^{\ell}\big((X, \mathbf{x}), (Y, \boldsymbol{\theta})\big)$:

- $g^{\ell}(\mathbf{x}, \boldsymbol{\theta}) = \langle \mathbf{x}, \boldsymbol{\theta} \rangle$:

$$\prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(m_d - m_d^{\ell})^2}{\sigma_d^2 + (\sigma_d^{\ell})^2}\right) \langle \mathbf{x}, \boldsymbol{\theta} \rangle, \tag{71}$$

- $k = k_{ntsk}^{\gamma}$ and $g^{\ell} = \langle \mathbf{x}, \boldsymbol{\theta} \rangle$, where $\mathbf{x} = (x_1, \ldots, x_D)$, $\boldsymbol{\theta} = (x_1^{sup}, \ldots, x_d^{sup}, \ldots, x_D^{sup})$, and x_d^{sup} given by (68):

$$\prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(m_d - m_d^{\ell})^2}{\sigma_d^2 + (\sigma_d^{\ell})^2 + \gamma}\right) \sum_{d=1}^{D} x_d \, \frac{m_d(\sigma_d^{\ell})^2 + m_d^{\ell}\sigma_d^2}{\sigma_d^2 + (\sigma_d^{\ell})^2}, \tag{72}$$

- g^{ℓ} is a RBF kernel:

$$\prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(m_d - m_d^{\ell})^2}{\sigma_d^2 + (\sigma_d^{\ell})^2}\right) \exp\Big(-\gamma\frac{1}{2}\|\mathbf{x}^I - \boldsymbol{\theta}^{\ell}\|^2\Big). \tag{73}$$

There are innumerable possibilities to engineer kernels on fuzzy sets from (67), by choosing different fuzzy sets and T-norms for kernel k_{ntsk}, and different kernels g^{ℓ}.

H. The non-singleton TSK kernel and fuzzy equivalence relations

Positive definite kernels generating matrices with a value one in the diagonal, and taking values on $[0, 1]$ are fuzzy equivalence relations with respect to a T-norm (Collorary 6 in [104]). Such kernels can be interpreted as fuzzy logic formulas for fuzzy rules (Theorem 9 in [105]). We show that non-singleton TSK kernels satisfy the two mentioned conditions.

Definition 18 (Fuzzy equivalence relation)**.** Mapping $\mathcal{E} : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is a fuzzy equivalence relation w.r.t. a T-norm T if:

1) $\forall x \in \mathcal{X},\ \mathcal{E}(x, x) = 1$;
2) $\forall x, y \in \mathcal{X},\ \mathcal{E}(x, y) = \mathcal{E}(y, x)$;
3) $\forall x, y, z \in \mathcal{X},\ T(\mathcal{E}(x, y), \mathcal{E}(y, z)) \leq \mathcal{E}(x, z)$.

are satisfied.

Lemma 8 (Kernels are at leas $T_{cos}transitivity$ [104])**.** *Let $\mathcal{X}$ be a non empty-set. Let $k : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be a positive definite kernel satisfying $\forall x \in \mathcal{X} : k(x, x) = 1$. Then $\forall x, y, z \in \mathcal{X}$, kernel k satisfy T_{cos} transitivity:*

$$T_{cos}(k(x, y), k(y, z)) \leq k(x, z), \tag{74}$$

where

$$T_{cos}(a, b) = \max(ab - \sqrt{1 - a^2}\sqrt{1 - b^2}, 0), \tag{75}$$

is a archimedean T-norm and it is the greatest T-norm with such property [104].

Lemma 9 (Kernels as fuzzy logic formulas for fuzzy rules [105])**.** *Let $\mathcal{X}$ be an non-empty set. Let $k : \mathcal{X} \times \mathcal{X} \to [0, 1]$ a positive definite kernel satisfying $\forall x \in \mathcal{X} : k(x, x) = 1$. Then $\forall x, y, z \in \mathcal{X}$, there is a class of fuzzy sets $\mu_{i\in I} : \mathcal{X} \to [0, 1]$, where I is a non-empty set, satisfying*

$$\forall x, y \in \mathcal{X} : k(x, y) = \inf_{i\in I} \overleftrightarrow{T}_M(\mu_i(x), \mu_i(y)), \tag{76}$$

where $\overleftrightarrow{T}_M = \min(\overrightarrow{T}(x, y), \overrightarrow{T}(y, x))$ is a bi-implication operator and $\overrightarrow{T}(x, y) = \sup\{t \in [0, 1] | T(x, t) \leq y\}$, is a implication function generated from a T-norm T [105].

We state the following

Lemma 10. *Non-singleton TSK kernel on fuzzy sets satisfy (Lemma* (8)*) and Lemma* (8) *if normal fuzzy sets were used.*

[17] Derivation is achieved using Lemma 6.

[18] See Equation (67)

Kernel Equation	SVs	C	γ	Acc.
-	11.8	2^{15}	2^{-4}	96
(70)	31	2^{1}	2^{-5}	**98**
(73)	37.1	2^{2}	2^{-4}	**98**
(71)*	41.5	2^{0}	-	**98**

$*$ It is a kernel without parameter.

Table IV
RESULTS ON THE DATASET *100mlI-4*.

I. Experiments on supervised classification over an athletics performance dataset

The dataset *100mlI-4* [120] contain information about athletics performance. This data is used to classify when a mark of 100 meters is achieved by some athlete. Each observation has the following four interval-valued features: the height vs weight rate, reaction time, velocity at 20 meters and maximum velocity at 40 meters. Each observation was imprecise because three different judges scored each athlete.

1) Scaling and fuzzyfication: We used a support vector machine with some kernels on fuzzy sets presented in this section. All the interval data was scaled to be in the interval unitary. Each interval value $|\underline{x}_d, \overline{x}_d|$ was modeled by a fuzzy set with a Gaussian membership function: $X_d : \mathbb{R} \to [0, 1]$, with parameters $m_d \in \mathbb{R}$ and $\sigma_d \in \mathbb{R}^+$ given by the mean of each interval and

$$\sigma_d = \frac{|\underline{x}_d - \overline{x}_d|}{2\sqrt{2}\ln 2}. \tag{77}$$

Then the interval-valued dataset:

$$\{[\underline{x}_1^{(i)}, \overline{x}_1^{(i)}], [\underline{x}_2^{(i)}, \overline{x}_2^{(i)}], \ldots, [\underline{x}_D^{(i)}, \overline{x}_D^{(i)}]\}_{i=1}^N,$$

is transformed into the fuzzy dataset:

$$\{(X_1^{(i)}, X_2^{(i)}, \ldots, X_D^{(i)})\}_{i=1}^N.$$

2) Model selection: We perform model selection over the regularization parameter C of SVM and the kernel parameter γ. We chose from the grid $C = \{2^{-1}, \ldots, 2^{14}\}$ vs. $\gamma = \{2^4, \ldots, 2^{-10}\}$ the pair (C, γ) with the best 10-fold cross validation accuracy. We solved ties by choosing the pair with less number of support vectors. For kernels without parameter, we only performed grid search over C. As this dataset is very small, we only report the cross validation accuracy, we believe that this value will reflect the test accuracy. The baseline kernel for comparison purposes was the Gaussian kernel k_γ^{rbf} over the mean values of the intervals.

3) Results: Table IV shows the kernels used in the experiment, the best regularization parameter C, the best kernel parameter γ, the cross validation accuracy (Acc.) and the mean of support vectors from the 10-fold cross validation for the pair (C, γ). The Gaussian kernel is denoted by *RBF*, and the non-singleton TSK kernel by TSK-NS.

Results suggest that TSK kernels have good performance in terms of accuracy and number of support vectors for this dataset.

J. Experiment on dyslexia prediction dataset

The dataset *Dyslexic-12-4* [68], [120], [121] contains data used to predict whether or not patient present dyslexia. All the data was collected from a graphical test called, Bender test. This data was preprocessed in order to include the subjectivity of the specialist's evaluation. This give an interval-valued dataset. The dataset has twelve features and four classes: {dyslexia, no dyslexia, control, other}.

Kernel Equation	SVs	C	γ	Acc.
-	33	2^{5}	2^{-8}	36.0
(70)	33.3	2^{3}	2^{5}	36.0
(71)*	45	2^{1}	-	**38.5**

$*$It is a kernel without parameter.

Table V
RESULTS ON THE DATASET *Dyslexic-12-4*.

1) Results: The data was scaled, fuzzyfied as the last experiment. The model selection was also the same of the last experiment. Table V shows the kernels and the results of this experiment. Results suggest that kernels on fuzzy sets have a good performance in terms of accuracy for this dataset. Notice that the kernel on fuzzy set with better accuracy has not kernel parameter, consequently, the training time for this kernel was faster than the others.

K. Summary and further research on this topic

We presented in this section how TSK fuzzy systems induce kernels. Moreover, we define an extension of this fuzzy system to the case of having fuzzy sets as inputs. We call this fuzzy system as the non-singleton fuzzy system. Based on this new fuzzy system, we define a new class of kernels on fuzzy sets: *the non-singleton TSS fuzzy kernel on fuzzy sets*. We show that this kernel is based on the intersection of fuzzy sets. We also prove when this kernel is positive definite. We give some examples of implementation of this kernel. We prove that this kernel is a fuzzy equivalence relation. Finally, we made an experimental study using a support vector machine with our kernels on two supervised classification tasks: athletics performance and dyslexia prediction. As further research, we will look for practical applications on other tasks. Also, we will investigate if other kinds of fuzzy systems can induce kernels.

VI. A DISTANCE-BASED KERNEL ON FUZZY SETS

Similarity measures between fuzzy sets is a hot topic of research because of its several applications in ML, computer vision, image processing, statistical tests, etc. In this section, we propose a *distance-based kernel on fuzzy sets*. Our idea is based on putting a distance between fuzzy sets into the kernel definition. If the distance is a metric, the distance-based kernel on fuzzy sets is a similarity measure on fuzzy sets with a geometrical interpretation in a RKHS; pseudo-metrics and

semi-metrics on fuzzy sets still can be used. We successfully used our proposed kernels on a hypothesis testing task on fuzzy data.

A. Contributions

Our main contributions are:

- the formulation of kernels on fuzzy sets by using distances between fuzzy sets,
- a new formulation of distances between fuzzy sets using kernels.
- a two-sample kernel hypothesis testing on fuzzy data using our kernels.

B. Distance induced by positive kernels

We use the term *distance* to denote indistinctly a metric, a pseudo-metric, or a semi-metric. Notice that the following is satisfied in a RKHS $\mathcal{H}$ with kernel k:

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

Not only positive definite kernels induce metrics in $\mathcal{H}$, but also *Conditionally Positive Definite* (CPD) kernels. It is still possible to relax the condition of positive definiteness for kernels by only requiring symmetric kernels to be used to define distances. Of course, such induced distances could violate the definition of metric, nevertheless they would induce zero-diagonal symmetric distances.

C. Distances on fuzzy sets induced by kernels

If k is a kernel on fuzzy sets. Distances between fuzzy sets could be formulated using kernels on fuzzy sets as follows:

$$D(X,Y) = k(X,Y) - 2k(X,Y) + k(Y,Y). \quad (78)$$

We call it as a *kernel-based distance on fuzzy sets*. A consequence of (78) in fuzzy set theory is that if k is positive definite, then Equation (78) give us a way to define 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 k.

D. A distance-based kernel on fuzzy sets

We present a new class of kernels on fuzzy sets that are distance substitution kernels [122]. The distance being substituted in the kernel definition is a distance between fuzzy sets. If D is a metric, the following kernels on fuzzy sets are positive definite:

$$K(X,Y) = \langle X,Y\rangle_D^{x_0} \quad (79)$$

$$K(X,Y) = \big(\alpha + \gamma\langle X,Y\rangle_D^{X_0}\big)^{\beta} \quad (80)$$

$$K(X,Y) = \exp(-\gamma D(X,Y)^2), \quad (81)$$

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

$$\langle X,Y\rangle_D^{X_0} \overset{\text{def}}{=} \frac{1}{2}\big(D(X,X_0)^2 + D(Y,X_0)^2 - D(X,Y)^2\big) \quad (82)$$

We call all of those kernels as *distance-based kernels on fuzzy sets*. Note that we can use pseudo-metrics or a semi-metrics on fuzzy sets D. Resulting kernels will not be positive definite, but they will have practical applications and a geometrical interpretation. [57], [58], [123], [124]. Some popular distances between fuzzy sets that could induce new kernels on fuzzy sets are given in Table VI. See [88], [125]–[127] and references therein for details.

E. Application to the task of two-sample kernel hypothesis testing

We present an experimental study using distance-based kernels on fuzzy sets on a hypothesis testing task. We use real clinical data of cancer prognosis. Some variables of this dataset are vague and imprecise, then they are better described by fuzzy sets. The problem to be solved is formulated as follows: Given two populations, determine whether they were drawn from the same distribution or not. In order to do that, we used the kernel two-sample test [45]. The novelty, is that we extend the method to the case of fuzzy data using kernels on fuzzy sets[19].

F. Kernel hypothesis testing

Kernels had been used to perform statistical tests to decide wheather two samples were drawn from the same or different distributions [45], [129]–[131]. In this experiment we used a kernel two-sample test approach. A main concept for this procedure is the Maximum Mean Discrepancy (MMD) [45].

Definition 19 (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 by:

$$\text{MMD}[\mathcal{F},\mathbb{P},\mathbb{Q}] = \sup_{f\in\mathcal{F}} \big(\mathbb{E}_{\text{X}\sim\mathbb{P}}[f(\text{X})] - \mathbb{E}_{\text{Y}\sim\mathbb{Q}}[f(\text{Y})]\big), \quad (83)$$

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

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 for MMD is given by (Lemma 6 from [45]):

$$\begin{aligned}\text{MMD}_u^2[\mathcal{F},s_X,s_Y] = {} & \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i,x_j) \\ & -\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j) \\ & +\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i,y_j) \end{aligned} \quad (84)$$

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: $H_A : \mathbb{P} \neq \mathbb{Q}$ holds. That is, if $\text{MMD}_u[\mathcal{F},s_X,s_Y] > \epsilon$, where ϵ is an arbitrary threshold, 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

[19]The code used in the experiments can 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
$\frac{\sum_{x\in\Omega} \lvert X(x) - X'(x)\rvert}{\sum_{x\in\Omega} \lvert X(x) + X'(x)\rvert}$	metric, discrete case
$1 - \frac{\sum_{x\in\Omega} \min[X(x), X'(x)]}{\sum_{x\in\Omega} \max[X(x), X'(x)]}$	semi-metric
$1 - \sum_{x\in\Omega} \frac{\min[X(x), X'(x)]}{\max[X(x), X'(x)]}$	semi-metric
$1 - \max_{x\in\Omega} \min[X(x), X'(x)]$	pseudo-metric
$\int_0^1 D(X_\alpha, X'_\alpha)d\alpha$	X_α is the α-cut set [128], and D is any metric on sets
$\frac{\sum_{x,x'\in\Omega} d(x,x') \min[X(x), X'(x')]}{\sum_{x,x'\in\Omega} \min[X(x), X'(x')]}$	metric

Table VI
DISTANCES ON FUZZY SETS THAT INDUCES KERNELS ON FUZZY SETS.

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. See Section 4 and 5 from [45] for details. In our experiments, we estimated the threshold using bootstrap.

G. Two-sample test with kernels on fuzzy sets

We show in this section the methodology used for the statistical test using kernels on fuzzy sets.

1) Breast cancer dataset: The *Breast Cancer* dataset [132][20], 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 VII. Note that the variables *age, menopause, tumor-size* and *inv-node* contain imprecise values.

2) Fuzzification of the dataset: Table VII shows that the dataset is heterogeneous, because it contains variables with precise and non-precise values. For the variables *age, menopause, tumor-size*, and *inv-nodes*, we used fuzzy sets to model the imprecise values. The fuzzification was performed as follows.

Variable *age* was modeled with fuzzy sets using trapezoidal membership functions given by:

$$X(x; a, b, c, d) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \frac{d-x}{d-c}, & c \le x \le d \\ 0, & d \le x \end{cases} \quad (85)$$

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

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 Table VII), 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 8 shows the trapezoidal fuzzy sets used to represent the values of the variable *age*.

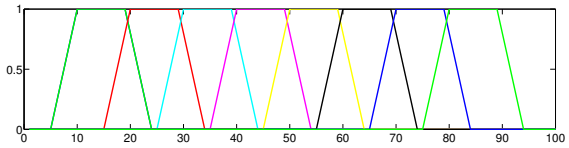


Figure 8. Fuzzification of variable *age* with trapezoidal fuzzy sets.

Variable *menopause*, was modeled by a fuzzy set with a *Z-shaped* membership function to represent the linguistic term *lq40* (less than forty years old), that is:

$$X(x; a, b) = \begin{cases} 1, & x \le a \\ 1 - 2\left(\frac{x-a}{b-a}\right)^2, & a \le x \le \frac{a+b}{2} \\ 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \le x \le b \\ 0, & x \ge b, \end{cases} \quad (86)$$

where $a = 40$ and $b = 45$. Figure 9 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:

$$X(x; a, b) = \begin{cases} 0, & x \le a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \le x \le \frac{a+b}{2} \\ 1 - 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \le x \le b \\ 1, & x \ge b, \end{cases} \quad (87)$$

where $a = 35$ and $b = 40$. Figure 9 shows this fuzzy set as a blue curve. To represent the linguistic value *premeno*, we used

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 VII
VARIABLES OF THE BREAST CANCER DATASET.

a fuzzy set with a Gaussian membership function, that is:

$$X(x;\sigma,c) = \exp\Big(-\frac{(x-c)^2}{2\sigma^2}\Big) \tag{88}$$

where $c = 45$ and σ was estimated to be $(50-40)/(2\sqrt{2\log 2})$[21]. Figure 9 show this fuzzy set as a yellow curve.

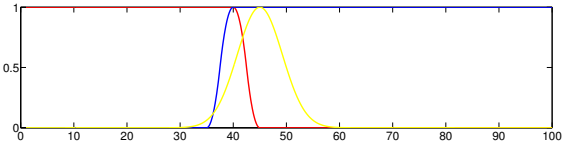


Figure 9. Fuzzification of the variable *menopause* with Gaussian (yellow), S-shaped (blue) and Z-shaped (red) fuzzy sets.

Variable *tumor-size* was 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 used to model the other values. Figure 10 shows the fuzzy sets used to model this variable.

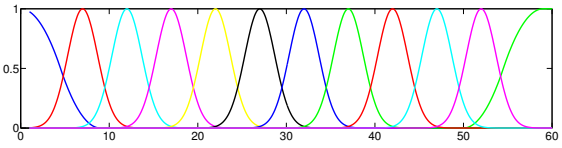


Figure 10. Fuzzification of the variable *tumor-size*, with S-shaped (leftmost), Z-shaped (rightmost) and Gaussian fuzzy sets.

Finally, Variable *inv-nodo* was modeled as 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 11 shows the fuzzy sets used to model this variable.

For comparison purposes, we constructed a crisp version of this dataset, considering the mean values of each interval.

3) Experiments: We performed two experiments. **Experiment I**: The purpose of this experiment is to test it the MMD can correctly identify whether or not two samples

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

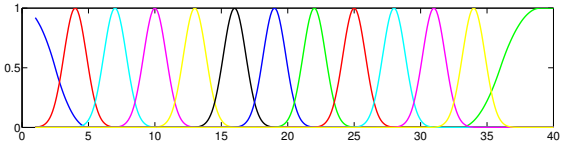


Figure 11. Fuzzification of the variable *inv-nodes* with with S-shaped (leftmost), Z-shaped (rightmost) and Gaussian fuzzy sets.

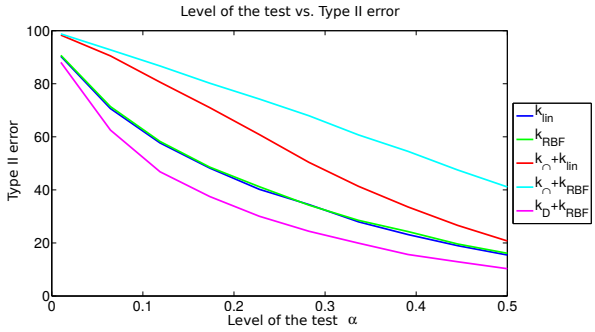


Figure 12. Two-sample test with five different kernels. The graph shows the results in terms of the Type II error, vs the evel of the test α.

were drawn from two *different* distributions. We randomly sampled 25 observations from patients showing breast cancer recurrence and 25 observations from patients without breast cancer recurrence to form a pair of samples to be tested. We set α to 0.05. as the level of the test. We applied the test over one hundred randomly chosen pairs of samples. Then we get the frequency the null hypothesis is rejected. In order to get statistics of mean and variance, we performed the above procedure 250 times.

Experiment II: We used the same procedure as before but this time a pair of samples was chosen from the *same* category (recorrence, or non recorrence). The category whereby the samples were chosen was randomly selected. We proceeded to get the mean and the variance as before.

Kernels for the crisp dataset We used the linear kernel: $k_{lin}(x,y) = x^T y,\ x,y \in \mathbb{R}^D$, and the RBF kernel: $k_{RBF}(x,y) = \exp(-\gamma\|x-y\|^2),\ x,y \in \mathbb{R}^D,\ \gamma \in \mathbb{R}^+$. We estimate the γ parameter using the median heuristic: we compute the median of the all possible distances between observations in the dataset and we set γ to be the one divided by this value.

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 VIII
RESULTS FOR $\alpha = 0.05$. THE NULL HYPOTHESIS IS DENOTED BY H_0 .

Kernels for the fuzzy dataset As the data is heterogeneous, we decided to use a convex combination of kernels:

$$k(x, y) = \sum_{i=1}^{K} \beta_i k_i(x_i, y_i), \tag{89}$$

where $\beta_i \geq 0$ and $\sum_i \beta_i = 1$. This kernel has some advantages as: different kernels can be used in different subsets of variables with values on different domains, and it is possible to optimize the β_i values from data. The resulting kernel is positive definite because any linear combination of kernels is also a positive definite kernel.

We denoted by $k_\cap + k_{lin}$, the kernel given by the convex combination of linear kernels on crisp variables and the intersection kernel on fuzzy sets over fuzzy variables (*age, menopause, tumor-size* and *inv-nodes*) [50]. We used the minimum T-norm operator for the intersection kernel on fuzzy sets[22].

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

Finally, we used a convex combination of a RBF kernel on crisp variables and a distance-based kernel on fuzzy sets with the following metric on fuzzy sets:

$$D(X, X') = \frac{\sum_{x\in\Omega} |X(x) - X'(x)|}{\sum_{x\in\Omega} |X(x) + X'(x)|}. \tag{90}$$

The distance-based kernel on fuzzy sets was given by:

$$K_D(X, X') = \exp(-\lambda D(X, X')^2) \tag{91}$$

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) Results: Figure 12 shows the performance of the kernels on graph of the Type II error vs. the level of the test α. We note that kernel $k_D + k_{RBF}$ has lower Type II error than the other ones, for several choices of α.

Table VIII shows the results for the two-sample tests for $\alpha = 0.05$. The first column shows the type of the experiment. The second column shows the condition of the samples. The word *accepted* in the third column 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 of Section VI-G3. 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 very similarly. The second row shows the case when the null hypothesis is rejected despite of being true. That is called a Type I error, this value is bounded by the level of the test.

The third row shows the Type II error: the test wrongly says that both samples were drawn from the same distribution. That is the test says that the alternative hypothesis is false despite of being true. Type II errors give us a way to measure the relevance of the kernels in this hypothesis testing task. We observed that the distance-based kernel on fuzzy sets $K_D + k_{RBF}$ is the best of all of them achieving lower Type II error for this dataset.

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. The kernels $k_\cap + k_{lin}$ and $k_\cap + k_{RBF}$ performs badly and the kernel $K_D + k_{RBF}$ has a higher value than the others.

[22]This kernel has not bandwidth parameter, then is not necessary to perform a parameter search.

H. Summary and further research on this topic

In this section, we proposed a new class of kernels on fuzzy sets: a *distance-based kernel* on fuzzy sets. Those kernels are based on the concept of distance substitution kernel [122]. A distance-based kernel on fuzzy sets is given by setting a distance on fuzzy sets within the kernel definition. As a further area of research, we investigate how to extend this concept for intuitionistic fuzzy sets [133] and hesitant fuzzy sets (See [133], [134]). [134].

VII. CONCLUSIONS AND FURTHER RESEARCH

In this research, we investigated how kernel methods, fuzzy sets and probability measures can be used to deal with datasets of the form of (1). We stated that, depending on the nature of s_i, it is possible to use either fuzzy sets or probability measures to model such observations. Using kernels, we used the kernel embedding of either probability measures or fuzzy sets to perform data analytics. As a result we defined a new class of kernels on fuzzy sets. Moreover, we defined the *support measure data description models* as a data description models for distributional data. Thus we proved that the methodology used in this research is a valid approach. As further step, we will extend those methods to other ML tasks over datasets of interest.

REFERENCES

[1] L. Xiong, B. Póczos, and J. G. Schneider, "Group anomaly detection using flexible genre models," in *NIPS*, 2011, pp. 1071–1079.

[2] L. Xiong, B. Póczos, J. G. Schneider, A. J. Connolly, and J. VanderPlas, "Hierarchical probabilistic models for group anomaly detection," in *AISTATS*, 2011, pp. 789–797.

[3] B. Póczos, L. Xiong, and J. G. Schneider, "Nonparametric divergence estimation with applications to machine learning on distributions," *CoRR*, vol. abs/1202.3758, 2012.

[4] K. Muandet and B. Schoelkopf, "One-class support measure machines for group anomaly detection," in *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*. Corvallis, Oregon: AUAI Press, 2013, pp. 449–458.

[5] J. Guevara, R. Hirata, and S. Canu, "Kernel functions in takagi-sugeno-kang fuzzy system with nonsingleton fuzzy input," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, 2013, pp. 1–8.

[6] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *J. Mach. Learn. Res.*, vol. 7, pp. 1283–1314, Dec. 2006.

[7] A. Ben-Tal, S. Bhadra, C. Bhattacharyya, and J. S. Nath, "Chance constrained uncertain classification via robust optimization," *Mathematical programming*, vol. 127, no. 1, pp. 145–173, 2011.

[8] J. B. T. Zhang, "Support vector classification with input data uncertainty," in *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, vol. 17. MIT Press, 2005, p. 161.

[9] R. Viertl, *Statistical Methods for Fuzzy Data*, ser. Wiley Series in Probability and Statistics. Wiley, 2011.

[10] T. Graepel and R. Herbrich, "Invariant pattern recognition by semidefinite programming machines," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2003, p. 2004.

[11] J. Yang and S. Gunn, "Exploiting uncertain data in support vector classification," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, B. Apolloni, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2007, vol. 4694, pp. 148–155.

[12] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space embedding for distributions," in *Algorithmic Learning Theory*. Springer, 2007, pp. 13–31.

[13] R. Kondor and T. Jebara, "A kernel between sets of vectors," in *ICML*, 2003, pp. 361–368.

[14] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Scholkopf, "Learning from distributions via support measure machines," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 10–18.

[15] G. Wahba, *Spline models for observational data*. Siam, 1990, no. 59.

[16] B. Schölkopf and A. J. Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*, ser. Adaptive computation and machine learning. MIT Press, 2002.

[17] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press, 2004.

[18] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Boston, 2004, vol. 3.

[19] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 209, pp. pp. 415–446, 1909. [Online]. Available: http://www.jstor.org/stable/91043

[20] A. N. Kolmogorov, "Stationary sequences in Hilbert space," *Bull. Math. Univ. Moscow*, vol. 2, no. 6, 1941.

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

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

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

[24] E. Pekalska, P. Paclik, and R. P. Duin, "A generalized kernel approach to dissimilarity-based classification," *The Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2002.

[25] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[26] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 144–152.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://dx.doi.org/10.1023/A:1022627411411

[28] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[29] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[30] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 775–782.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[33] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[34] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, and S. García, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework." *Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.

[35] R. Viertl, *Statistical Methods for Fuzzy Data*, ser. Wiley Series in Probability and Statistics. Wiley, 2011.

[36] J. M. Mendel, *Uncertain rule-based fuzzy logic system: introduction and new directions*. Prentice–Hall PTR, 2001.

[37] E. Hüllermeier, "Does machine learning need fuzzy logic?" *Fuzzy Sets and Systems*, vol. 281, pp. 292–299, 2015.

[38] L. Wernisch, S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. D. Butcher, and N. G. Stoker, "Analysis of whole-genome microarray replicates using mixed models," *Bioinformatics*, vol. 19, no. 1, pp. 53–61, 2003.

[39] S. Bhadra, "Learning robust support vector machine classifiers with uncertain observations," Ph.D. dissertation, 2012.

[40] S. Bhadra, C. Bhattacharyya, N. R. Chandra, and I. S. Mian, "A linear programming approach for estimating the structure of a sparse linear genetic network from transcript profiling data," *Algorithms for Molecular Biology*, vol. 4, no. 1, p. 5, 2009.

[41] S. Bhadra, J. S. Nath, A. Ben-Tal, and C. Bhattacharyya, "Interval data classification under partial information: A chance-constraint approach," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2009, pp. 208–219.

[42] S. Bhadra, S. Bhattacharya, C. Bhattacharyya, and A. Ben-Tal, "Robust formulations for handling uncertainty in kernel matrices," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 71–78.

[43] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *The Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[44] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *J. Mach. Learn. Res.*, vol. 6, pp. 2075–2129, Dec. 2005.

[45] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[46] L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola, "Hilbert space embeddings of hidden markov models," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 991–998.

[47] K. Fukumizu, L. Song, and A. Gretton, "Kernel bayes' rule," *arXiv preprint arXiv:1009.5736*, 2010.

[48] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, vol. 99, pp. 1517–1561, 2010.

[49] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift."

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

[51] ——, *Handbook of Fuzzy Set Comparisons - Theory and Applications*. Science Gate Publishing, 2015, ch. Fuzzy Set Similarity using a Distance-Based Kernel on Fuzzy Sets.

[52] J. Guevara, , and R. Canu, Stephane Hirata, "Support measure data description," Institute of Mathematics and Statistics, submitted to Pattern Analysis and Machine Intelligence, 2014.

[53] J. Guevara, R. Hirata, and S. Canu, "Support vector data description for uncertainty datasets," 2013, poster presented at The Machine Learning Summer School 26 August to 6 September 2013 at the Max Planck Institute for Intelligent Systems, Tübingen, Germany.

[54] J. Guevara, , R. Canu, Stephane Hirata, J. Guevara, R. Hirata, and S. Canu, "Support measure data description for group anomaly detection," *SIGACM KDD, ODDx3'15*, pp. 24–33, August 2015.

[55] J. Guevara, "Modelos de aprendizado supervisionado usando métodos kernel, conjuntos fuzzy e medidas de probabilidade," Ph.D. dissertation, Instituto de Matemática e Estatística, Universidade de Sao Paulo, 2015.

[56] B. Haasdonk, "Feature space interpretation of svms with indefinite kernels," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 4, pp. 482–492, 2005.

[57] B. Haasdonk and D. Keysers, "Tangent distance kernels for support vector machines," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2. IEEE, 2002, pp. 864–868.

[58] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in neural information processing systems*, 2003, p. None.

[59] H. S. K.-i. Noma, "Dynamic time-alignment kernel in support vector machine," *Advances in neural information processing systems*, vol. 14, p. 921, 2002.

[60] B. Scholkopf, "The kernel trick for distances," *Advances in neural information processing systems*, pp. 301–307, 2001.

[61] F. Yu and M. Zhang, "Generalized triangular norms based product and similarity of fuzzy sets." in *FSKD (1)'08*, 2008, pp. 286–290.

[62] E. P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*, 1st ed. Springer, 2000.

[63] M. L. Puri and D. A. Ralescu, "Fuzzy random variables," *Journal of mathematical analysis and applications*, vol. 114, no. 2, pp. 409–422, 1986.

[64] H. Kwakernaak, "Fuzzy random variables—i. definitions and theorems," *Information Sciences*, vol. 15, no. 1, pp. 1–29, 1978.

[65] A. F. Shapiro, "Fuzzy random variables," *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 307–314, 2009.

[66] A. Palacios and J. Alcala-Fdez, "Mining fuzzy association rules from low quality data," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, pp. 0–0, 2012.

[67] A. Palacios, L. Sánchez, and I. Couso, "Diagnosis of dyslexia with low quality data with genetic fuzzy systems," *International Journal of Approximate Reasoning*, vol. 51, no. 8, pp. 993–1009, 2010.

[68] A. M. Palacios, L. Sánchez, and I. Couso, "Future performance modeling in athletism with low quality data-based genetic fuzzy systems," *Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 207–228, 2011.

[69] L. Sánchez, I. Couso, and J. Casillas, "Genetic learning of fuzzy rules based on low quality data," *Fuzzy Sets and Systems*, vol. 160, no. 17, pp. 2524–2552, 2009.

[70] L. Zadeh, "Soft computing and fuzzy logic," *Software, IEEE*, vol. 11, no. 6, pp. 48 –56, nov. 1994.

[71] P. P. Wang, *Computing with words*. John Wiley & Sons, Inc., 2001.

[72] L. A. Zadeh, "Fuzzy logic= computing with words," *Fuzzy Systems, IEEE Transactions on*, vol. 4, no. 2, pp. 103–111, 1996.

[73] W. Polonik, "Minimum volume sets and generalized quantile processes," *Stochastic Processes and their Applications*, vol. 69, no. 1, pp. 1 – 24, 1997.

[74] C. Scott and R. D. Nowak, "Learning minimum volume sets." *Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[75] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[76] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[77] C. Guilbart, "Produits scalaires sur l'espace des mesures," in *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, vol. 15, no. 4. Gauthier-Villars, 1979, pp. 333–354.

[78] C. Suquet *et al.*, "Distances euclidiennes sur les mesures signees et applications a des theoremes de berry-esseen." *Bulletin of the Belgian Mathematical Society Simon Stevin*, vol. 2, no. 2, pp. 161–182, 1995.

[79] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 489–496.

[80] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective hilbert space embeddings of probability measures," in *In COLT*, 2008.

[81] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. Mach. Learn. Res.*, vol. 2, pp. 299–312, Mar. 2002.

[82] L. Debnath and P. Mikusiński, *Hilbert Spaces with Applications*. Elsevier Academic Press, 2005.

[83] C. A. Floudas and P. M. Pardalos, *Encyclopedia of optimization*. Springer Science & Business Media, 2008, vol. 1.

[84] P. Chan and M. Mahoney, "Modeling multiple time series for anomaly detection," in *Data Mining, Fifth IEEE International Conference on*, Nov 2005, pp. 8 pp.–.

[85] E. Keogh, J. Lin, and A. Fu, "Hot sax: efficiently finding the most unusual time series subsequence," in *Data Mining, Fifth IEEE International Conference on*, Nov 2005, pp. 8 pp.–.

[86] K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 169–176. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401915

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

[88] I. Bloch, "On fuzzy distances and their use in image processing under imprecision," *Pattern Recognition*, vol. 32, no. 11, pp. 1873–1895, 1999.

[89] Y. Kanzawa, Y. Endo, and S. Miyamoto, "On kernel fuzzy c-means for data with tolerance using explicit mapping for kernel data analysis," in *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1–6.

[90] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Multiple kernel fuzzy clustering," *Fuzzy Systems, IEEE Transactions on*, vol. 20, no. 1, pp. 120 –134, feb. 2012.

[91] M. Raza and F.-H. Rhee, "Interval type-2 approach to kernel possibilistic c-means clustering," in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, june 2012, pp. 1 –7.

[92] X. Yang, G. Zhang, J. Lu, and J. Ma, "A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises," *Fuzzy Systems, IEEE Transactions on*, vol. 19, no. 1, pp. 105 –115, feb. 2011.

[93] J. Wang, J. Hua, and J. Guo, "Fuzzy maximum scatter discriminant analysis with kernel methods," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, vol. 2, aug. 2010, pp. 560 –564.

[94] G. Heo and P. Gader, "Robust kernel discriminant analysis using fuzzy memberships," *Pattern Recogn.*, vol. 44, no. 3, pp. 716–723, Mar. 2011.

[95] Y. Chen and J. Wang, "Support vector learning for fuzzy rule-based classification systems," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 716 – 728, dec. 2003.

[96] Y. Chen, "Support vector machines and fuzzy systems," in *Soft Computing for Knowledge Discovery and Data Mining*, O. Maimon and L. Rokach, Eds. Springer US, 2008, pp. 205–223.

[97] J.-T. Jeng and T.-T. Lee, "Support vector machines for the fuzzy neural networks," in *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, vol. 6, 1999, pp. 115 –120 vol.6.

[98] C.-T. Lin, C.-M. Yeh, S.-F. Liang, J.-F. Chung, and N. Kumar, "Support-vector-based fuzzy neural network for pattern classification," *Fuzzy Systems, IEEE Transactions on*, vol. 14, no. 1, pp. 31 – 41, feb. 2006.

[99] C.-F. Juang, S.-H. Chiu, and S.-W. Chang, "A self-organizing ts-type fuzzy network with support vector learning and its application to classification problems," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 5, pp. 998 –1008, oct. 2007.

[100] C. T. Wee and T. W. Wan, "Efsvmrobus-fcm: Evolutionary fuzzy rule-based support vector machines classifier with fcm clustering," in *Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on*, june 2008, pp. 606 –612.

[101] S.-M. Zhou and J. Gan, "Constructing l2-svm-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 3, pp. 398 –409, june 2007.

[102] W.-Y. Cheng and C.-F. Juang, "An incremental support vector machine-trained ts-type fuzzy system for online classification problems," *Fuzzy Sets Syst.*, vol. 163, pp. 24–44, January 2011.

[103] J. M. Lski, "On support vector regression machines with linguistic interpretation of the kernel matrix," *Fuzzy Sets Syst.*, vol. 157, no. 8, pp. 1092–1113, Apr. 2006.

[104] B. Moser, "On the t-transitivity of kernels," *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1787 – 1796, 2006.

[105] ——, "On representing and generating kernels by fuzzy equivalence relations," *J. Mach. Learn. Res.*, vol. 7, pp. 2603–2620, Dec. 2006.

[106] F. Liu and X. Xue, "Design of natural classification kernels using prior knowledge," *Trans. Fuz Sys.*, vol. 20, no. 1, pp. 135–152, Feb. 2012.

[107] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels." in *ICML*, vol. 2, 2002, pp. 179–186.

[108] D. Haussler, "Convolution kernels on discrete structures," Technical report, Department of Computer Science, University of California at Santa Cruz, Tech. Rep., 1999.

[109] R. L. Schilling, *Measures, integrals and martingales*. Cambridge University Press, 2005, vol. 13.

[110] D. Haussler, "Convolution kernels on discrete structures," Technical report, Department of Computer Science, University of California at Santa Cruz, Tech. Rep., 1999.

[111] R. B. Ash and C. Doleans-Dade, *Probability and measure theory*. Academic Press, 2000.

[112] T. Gartner, *Kernels for structured data*. World Scientific, 2008, vol. 72.

[113] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels," *Neural networks for signal processing IX*, 1999.

[114] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," *Journal of Computational and Graphical Statistics*, 2012.

[115] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[116] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[117] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions On Systems Man And Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[118] L.-X. Wang and J. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least-squares learning," *Neural Networks, IEEE Transactions on*, vol. 3, no. 5, pp. 807 –814, sep 1992.

[119] J. M. Mendel, *Uncertain rule-based fuzzy logic system: introduction and new directions*. Prentice–Hall PTR, 2001.

[120] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, and S. García, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework." *Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.

[121] A. Palacios, L. Sánchez, and I. Couso, "Extending a simple genetic cooperative-competitive learning fuzzy classifier to low quality datasets," *Evolutionary Intelligence*, vol. 2, no. 1-2, pp. 73–84, 2009.

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

[123] 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*, 2002, pp. 49–54.

[124] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1055–1064, 1999.

[125] A. Rosenfeld, "Distances between fuzzy sets," *Pattern Recognition Letters*, vol. 3, no. 4, pp. 229–233, 1985.

[126] B. Chaudhur and A. Rosenfeld, "On a metric distance between fuzzy sets," *Pattern Recognition Letters*, vol. 17, no. 11, pp. 1157–1160, 1996.

[127] P. Diamond, P. E. Kloeden, P. E. Kloeden, A. Mathematician, and P. E. Kloeden, *Metric spaces of fuzzy sets: theory and applications*. World Scientific, 1994.

[128] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic*. Prentice hall New Jersey, 1995, vol. 4.

[129] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[130] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Advances in neural information processing systems*, 2006, pp. 513–520.

[131] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, "A fast, consistent kernel two-sample test," in *Advances in neural information processing systems*, 2009, pp. 673–681.

[132] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[133] G. A. Papakostas, A. G. Hatzimichailidis, and V. G. Kaburlasos, "Distance and similarity measures between intuitionistic fuzzy sets: A comparative analysis from a pattern recognition point of view," *Pattern Recognition Letters*, vol. 34, no. 14, pp. 1609–1622, 2013.

[134] Z. Xu and M. Xia, "Distance and similarity measures for hesitant fuzzy sets," *Information Sciences*, vol. 181, no. 11, pp. 2128–2138, 2011.