Generalized and Customizable Sets in R

David Meyer
WU Wirtschaftsuniversität Wien

Kurt Hornik
WU Wirtschaftsuniversität Wien

Abstract

This introduction to the R package **sets** is a (slightly) modified version of Meyer and Hornik (2009a), published in the *Journal of Statistical Software*.

We present data structures and algorithms for sets and some generalizations thereof (fuzzy sets, multisets, and fuzzy multisets) available for R through the **sets** package. Fuzzy (multi-)sets are based on dynamically bound fuzzy logic families. Further extensions include user-definable iterators and matching functions.

Keywords: R, set, fuzzy logic, multiset, fuzzy set.

1. Introduction

Only few will deny the importance of sets and set theory, building the fundamentals of modern mathematics. For theory-building typically axiomatic approaches (e.g., Zermelo 1908; Fraenkel 1922) are used. However, even the primal, "naive" concept of sets representing "collections of distinct objects" (Cantor 1895) discarding order and count information seems both natural and practical. The main operation being "is-element-of", sets alone are of limited *practical* use—they most of the times serve as basic building blocks for more complex structures such as relations and generalized sets. A common way is to consider pairs (X, m) with set X ("universe") and membership function $m : X \to D$ mapping each member to its "grade". The subset of X of elements with non-zero membership is called "support". In *multisets*, elements may appear more than once, i.e., $D = \mathbb{N}_0$ (m is also called the multiplicity function). There are many applications in computer science and other disciplines (for a survey, see, e.g., Singh, Ibrahim, Yohanna, and Singh 2007). In statistics, multisets appear as frequency tables. *Fuzzy sets* have become quite popular since their introduction by Zadeh (1965). Here, the membership function maps into the unit interval. An interesting characteristic of fuzzy sets is that the actual behavior of set operations depends on the underlying fuzzy logic employed, which can be chosen according to domain-specific needs. Fuzzy sets are actively used in fields such as machine learning, engineering, medical science, and artificial intelligence (Dubois, Prade, and Yager 1996). *Fuzzy multisets* (Yager 1986) combine both approaches by allowing each element to map to more than one fuzzy membership grade, i.e., D is the power set of multisets over the unit interval. Examples for the application of fuzzy multisets can be found in the field of information retrieval (e.g., Matthé, Caluwe, de Tré, Hallez, Verstraete, Leman, Cornelis, Moelants, and Gansemans 2006).

The use of sets and variants thereof is common in modern general purpose programming languages: Java and C++ provide corresponding abstract data types (ADTs) in their class libraries, Pascal and Python offer sets as native data type. Indeed, since set elements are order-

invariant and unique, lookup mechanisms can be implemented very efficiently (for example via hashing, resulting in nearly constant run-time complexity, compared to linear search, requiring $n/2$ steps on the average for n elements). Surprisingly enough, sets are not standard in many mathematical programming environments such as MATLAB and Mathematica, and also R. Although the two latter offer set operations such as union and intersection, these are applied to linearly indexable structures (lists and vectors, respectively), *interpreting* them as sets. When it comes to R, this emulation is far from complete, and occasionally leads to inconsistent behavior. First of all, the existing infrastructure has no clear concept of how to compare elements, leading to possibly confusing results when different data types are involved in computations:

```
R> s <- list(1, "1")
R> union(s, s)

[[1]]
[1] 1

[[2]]
[1] "1"

R> intersect(s, s)

[[1]]
[1] 1
```

The reason is that most of the existing operations rely on `match()` which automatically performs type conversions disturbing in this context. Also, quite a few other basic operations such as the Cartesian product, the power set, the subset predicate, etc., are missing, let alone more specialized operations such as the closure under union or intersection. Then, the current facilities do not make use of a class system, making extensions hard (if not impossible). Another consequence is that no distinction can be made between sequences (ordered collections of objects) and sets (unordered collections of objects), which is key for the definition of complex structures where both concepts are combined such as relations. Also, there is no support in base R for extensions such as fuzzy sets or multisets.

A few extension packages available from the Comprehensive R Archive Network deal with fuzzy concepts: Package **fuzzyFDR** (Lewin 2007) calculates fuzzy decision rules for multiple testing, but does not provide any explicit data structures for fuzzy sets. The main functions in **fso** (Roberts 2007) for fuzzy set ordination compute and return, among other information, membership values represented by numeric matrices for some variables of the the input data. **fuzzyRankTests** (Geyer 2007) provides statistical tests based on fuzzy p values and fuzzy confidence intervals, the latter being returned as two separate numeric vectors for values and memberships. The **gcl** package (Vinterbo 2007) infers fuzzy rules from the input data, encapsulated in a classifying function returned by the training function. The rules are composed of triangular fuzzy sets, represented by triples describing the triangles' corner points for which the memberships become $(0, 1, 0)$, respectively. Similarly, the **FKBL** package for fuzzy knowledge base learning (Alvarez 2007) uses sequences of triangular fuzzy sets, defined by a vector

of corner points. Finally, **fuzzyOP** (Aklan, Altindas, Macit, Umar, and Unal 2008) provides support for fuzzy numbers: A set of n numbers is represented by a $k \times 2n$ numeric matrix, where two consecutive columns represent (at most k) supporting points and memberships, respectively, of the corresponding piecewise linear membership function. If some numbers have fewer supporting points than others, the remaining cells are filled with missing values (`NA`s).

The **sets** package (Meyer and Hornik 2009b) presented here provides a flexible and customizable basic infrastructure for *finite* sets and the generalizations mentioned above, including basic operations for fuzzy logic. Apart from complementing the data structures implemented in base R, extension packages like the ones mentioned above could gain in flexibility from building on a common infrastructure, facilitating data exchange and leveraging synergies.

The remainder of the paper is structured as follows. In Section 2, we discuss the design rationale of data structures and core algorithms. Section 3 introduces the most important set operations. Section 4 starts with constructors and specific methods for generalized sets, followed by a more focused presentation of the fuzzy logic infrastructure, and of functionality for handling and visualizing membership information. Section 5 shows how generalized sets can further be customized by specifying user-definable matching functions and iterators. Section 6 presents three examples before Section 7 concludes.

2. Design issues

There are many ways of implementing sets. Choice and efficiency largely depend on the domain range (i.e., the number of possible values for each element). If the domain is relatively small, i.e. in the range of integral data types such as `byte`, `integer`, `word` etc., the probably most efficient representation is an array of bits representing the domain elements like in Pascal (Wirth 1983). Operations such as union and intersection can then straightforwardly be implemented using logical `OR` and `AND`, respectively. This approach obviously fails for intractably large domains (e.g., strings or recursive objects). Without further application knowledge, one needs to resort to generic container ADTs with efficient element access such as hash tables or search trees (for unique elements). Operations can then be implemented following the classical element-based definitions: Union by inserting all elements of the smaller set into the larger one; intersection by creating a new set with all elements of the smaller set also contained in the larger one; etc.

Clearly, set comparison must be permutation invariant. Some care is needed for nested sets. Assume, e.g., the comparison of $A = \{1, \{2, 3\}\}$ and $B = \{1, \{3, 2\}\}$ which clearly are identical. To implement set equality, a matching operator would be used to check if all elements of A are contained in B. If elements were internally stored in this order during creation, the objects representing $\{2, 3\}$ and $\{3, 2\}$ would be different. Comparing two set elements for equality would thus require to recursively compare all elements down the nested structures, which can quickly become infeasible computationally. We avoid this by using a canonical ordering during set creation, guaranteeing that identical sets have identical physical representation as well. We chose to sort elements using the natural order for numeric values, the Unicode character representation for strings, and the serialization byte sequence (as strings) for other objects. Eventually, the ordered elements are stored in a list.

For the **sets** package, further limitations are imposed by the extensions presented in Sections 4

and 5: Generalized sets require, for each element, the membership information, and we also support user-defined, high-level matching functions for comparing elements. Since operations defined for generalized sets basically operate on the memberships, it seems appropriate to store these as (generic) vectors in the same order than the corresponding elements. Thus, memberships of separate sets can simply be combined element-wise.

Many operations (e.g., testing for equality, subsetting, intersection, etc.) are based on matching elements of the sets involved. This is implemented by inserting the elements of the larger one into a hash table (we use hashed environments), and to look up the elements of the smaller set in this table (Knuth 1973, p. 391). As hash key, we use the elements' character representation. Since different objects might map to the same hash key, we actually store the *indexes* of the list elements, and match the actual objects using a simple linear search. (Note that since the element list is sorted, elements with same representation are grouped, so the search will typically be fast.)

The implementation is based on R's S3 class system, allowing the definition of generic functions, dispatching appropriate methods depending on the first argument's class information. Objects for sets, generalized sets, and customizable sets have classes '`set`', '`gset`', and '`cset`', respectively, with '`set`' inheriting from '`gset`' in turn inheriting from '`cset`'. Suitable operators (such as `&` for intersection) are then "overloaded" to dispatch the right internal function corresponding to the operands' classes by defining corresponding methods for group generics. Additionally, all operations can directly be accessed using the corresponding name combined with a `set_`, `gset_`, or `cset_` prefix to give the user the choice of up- or downcasts when objects of different class levels are involved in one computation. For example, consider the union of the set $\{1\}$ and the fuzzy set $\{2/0.5\}$: using the generic operator will give an error since the operands' classes differ. The user needs, in fact, resolve the semantic ambiguity by explicitly choosing the intended operation: If the result should be a generalized (fuzzy) set, `gset_union()` should be used. To make the result a set (stripping membership information), s/he employs `set_union()` instead.

3. Sets

The basic constructor for creating sets is the `set()` function accepting any number of R objects as arguments.

```
R> s <- set(1L, 2L, 3L)
R> print(s)

{1L, 2L, 3L}
```

For elements that are not sets or atomic vectors of length 1, the print method for sets will use labels indicating the class (and length for vectors):

```
R> set("test", c, set("a", 2.5), list(1, 2))

{"test", <<function>>, {"a", 2.5}, <<list(2)>>}
```

Mainly for cosmetic reasons, there is also a tuple class that can be used for vectors:

```
R> set(1, pair(1,2), tuple(1L, 2L, 3L))

{1, (1, 2), (1L, 2L, 3L)}
```

In addition, there is a generic `as.set()` function coercing suitable objects to sets.

```
R> s2 <- as.set(2:4)
R> print(s2)

{2L, 3L, 4L}
```

There are some basic predicate functions (and corresponding generic operators) defined for the (in)equality (`!=`, `==`), (proper) subset (`<`, `<=`), (proper) superset (`>`, `>=`), and element-of (`%e%`) operations:

```
R> set_is_empty(set())

[1] TRUE

R> set(1) <= set(1,2)

[1] TRUE
```

Note that all predicate functions are vectorized for convenience: 1:4 The sequence `1:4` as *one* element would be looked up by using `list(1:4)` on the left-hand side. The class-specific functions dispatched by the generic operators are `set_contains_element()`, `set_is_equal()`, etc. Other than these predicate functions, one can use `length()` for the cardinality:

```
R> length(s)

[1] 3
```

`c()` and `|` for the union, `&` for the intersection, `%D%` for the symmetric difference:

```
R> s | set("a")

{"a", 1L, 2L, 3L}

R> s & s2

{2L, 3L}

R> s %D% s2

{1L, 4L}
```

`*` and `^`n for the (n-fold) Cartesian product (yielding a set of n-tuples):

```
R> s * s2
```

```
{(1L, 2L), (1L, 3L), (1L, 4L), (2L, 2L), (2L, 3L), (2L, 4L), (3L,
 2L), (3L, 3L), (3L, 4L)}
```

```
R> s ^ 2L
```

```
{(1L, 1L), (1L, 2L), (1L, 3L), (2L, 1L), (2L, 2L), (2L, 3L), (3L,
 1L), (3L, 2L), (3L, 3L)}
```

and `2^` for the power set:

```
R> 2 ^ s
```

```
{{}, {1L}, {2L}, {3L}, {1L, 2L}, {1L, 3L}, {2L, 3L}, {1L, 2L, 3L}}
```

The class-specific functions `set_union()`, `set_intersection()`, and `set_symdiff()` accept more than two arguments.[1] It is also possible to compute the *relative* complement of a set X in Y, basically removing the elements of X from Y:

```
R> set_complement(set(1), set(1,2,3))
```

```
{2, 3}
```

Note, however, that for sets (as opposed to *generalized* sets), the concept of a "universe" is not necessarily required, and therefore the *absolute* complement of a set not a well-defined operation. In the **sets** package, objects of class '`set`' are special cases of generalized sets. To stay faithful to the simplicity of the original set concept, we *define* a '`set`' object's universe to be identical to the set itself. The *absolute* complement of a '`set`' object is therefore always the empty set:

```
R> !set(1)
```

```
{}
```

`set_combn()` returns the set of all subsets of specified length:

```
R> ## subsets
R> set_combn(s, 2L)
```

```
{{1L, 2L}, {1L, 3L}, {2L, 3L}}
```

[1] The n-ary symmetric difference of a collection of sets consists of all elements contained in an odd number of the sets in the collection.

`closure()` and `reduction()` compute the closure and reduction under union or intersection for a set *family* (i.e., a set of sets):

```
R> cl <- closure(set(set(1), set(2), set(3)), "union")
R> print(cl)

{{1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3}, {1, 2, 3}}

R> reduction(cl, "union")

{{1}, {2}, {3}}
```

The `Summary()` group methods will also work if defined for the elements:

```
R> sum(s)

[1] 6

R> range(s)

[1] 1 3
```

Because set elements are unordered, it is not allowed to use positional subscripting. However, sets can be subset and elements be replaced by using the elements as index themselves:

```
R> s2 <- set(1,2, c, list(1,2))
R> print(s2)

{<<function>>, 1, 2, <<list(2)>>}

R> s2[[c]] <- "foo"
R> s2[[list(1, 2)]] <- "bar"
R> print(s2)

{"bar", "foo", 1, 2}

R> s2[list("foo", 1)]

{"foo", 1}
```

Further, iterations over *all* elements can be carried out using `for()` and `lapply()`/`sapply()`:

```
R> sapply(s, sqrt)

[1] 1.000000 1.414214 1.732051
```

```
R> for (i in s) print(i)

[1] 1
[1] 2
[1] 3
```

Note that `for()` only works because the underlying C code ignores the class information, and directly processes the low-level list representation instead. This will be replaced by a more intelligent "foreach" mechanism as soon as it exists in base R. `sapply()` and `lapply()` call the generic `as.list()` function before iterating over the elements. Since a corresponding method exists for sets objects, this is "safer" than using `for()`.

Using `set_outer()`, it is possible to apply a function on all factorial combinations of the elements of two sets. If only one set is specified, the function is applied on all pairs of this set.

```
R> set_outer(set(1,2), set(1,2,3), "/")

  1   2         3
1 1 0.5 0.3333333
2 2 1.0 0.6666667
```

4. Generalized sets

There are several extensions of sets such as *fuzzy sets* and *multisets*. Both can be be seen as special cases of *fuzzy multisets*. All have in common that they are defined on some *universe* of elements, and that each element maps to some membership information. We present how generalized sets are constructed, and demonstrate the effect of choosing different fuzzy logic families.

4.1. Constructors and specific methods

Generalized sets are created using the `gset()` function. The required arguments depend on whether membership information is specified extensionally (listing members) or intensionally (giving a rule for membership):

1. Extensional specification:
   a) Specify support and memberships as separate vectors. If memberships are omitted, they are assumed to be 1.
   b) Specify a set of elements along with their individual membership grades, using the element function (`e()`).
2. Intensional specification: Specify universe and membership function.

Note that for efficiency reasons, `gset()` will not store elements with zero memberships grades, and the specification of a universe is only required with membership functions. For convenience (and storage efficiency), a default universe can be defined using `sets_options()`. Set-specific universes supersede the default universe, if any. If no universe (general or specific) is defined, the support of a set will be interpreted as its universe. For multisets, the definition of a (general or set-specific) universe can be complemented by a maximum multiplicity or *bound*.

Without membership information, `gset()` creates a set (the support is converted to a set internally):

```
R> X <- c("A", "B", "C")
R> gset(support = X)

{"A", "B", "C"}
```

Note, however, that unlike for '`set`' objects, it is possible to define a universe that differs from (i.e., is a proper superset of) the support:

```
R> gset(support = X, universe = LETTERS[1:10])

{"A", "B", "C"}
```

A multiset requires an integer membership vector:

```
R> multi <- 1:3
R> gset(support = X, memberships = multi)

{"A" [1], "B" [2], "C" [3]}
```

For fuzzy sets, the memberships need to be out of the unit interval:

```
R> ms <- c(0.1, 0.3, 1)
R> gset(support = X, memberships = ms)

{"A" [0.1], "B" [0.3], "C" [1]}
```

Alternatively to separate support/membership specification, each elements can be paired with its membership value using `e()`:

```
R> gset(elements = list(e("A", 0.1), e("B", 0.2), e("C", 0.3)))

{"A" [0.1], "B" [0.2], "C" [0.3]}
```

Fuzzy sets can, additionally, be created using a membership function, applied to a specified (or the default) universe:

```
R> f <- function(x) switch(x, A = 0.1, B = 0.2, C = 1, 0)
R> gset(universe = X, charfun = f)

{"A" [0.1], "B" [0.2], "C" [1]}
```

For fuzzy multisets, the membership argument expects a list of membership grades, either specified as vectors, or as multisets:

```
R> ms2 <- list(c(0.1, 0.3, 0.4), c(1, 1),
+    gset(support = ms, memberships = multi))
R> gset(support = X, memberships = ms2)

{"A" [{0.1, 0.3, 0.4}], "B" [{1 [2]}], "C" [{0.1 [1], 0.3 [2], 1
 [3]}]}
```

`gset_cardinality()` returns the (relative) cardinality of a generalized set, computed as the sum (mean) of all memberships. `gset_support()`, `gset_memberships()`, `gset_height()` and `gset_core()` can be used to retrieve support, memberships, height (maximum membership degree), and the core (elements with membership 1), respectively, of a generalized set. `gset_charfun()` returns a (point-wise defined) characteristic function for a given gset. Note that in general, this will be different from the characteristic function possibly used for the creation.

As for sets, the usual operations such as union and intersection are available:

```
R> X <- gset(c("A", "B", "C"), 4:6)
R> Y <- gset(c("B", "C", "D"), 1:3)
R> X | Y

{"A" [4], "B" [5], "C" [6], "D" [3]}

R> X & Y

{"B" [1], "C" [2]}
```

Additionally, the product (`gset_product()`), sum (`+`), and difference (`-`) of sets are defined, which multiply, add, and subtract multiplicities (or memberships for fuzzy sets):

```
R> X + Y

{"A" [4], "B" [6], "C" [8], "D" [3]}

R> X - Y

{"A" [4], "B" [4], "C" [4]}

R> gset_product(X, Y)
```

```
{"B" [5], "C" [12]}
```

For fuzzy (multi-)sets, not only the relative, but also the absolute complement (`!`) is defined:

```
R> !gset(1, 0.3)

{1 [0.7]}

R> X <- gset("a", universe = letters[1:3])
R> !X

{"b", "c"}

R> !!X

{"a"}

R> !gset(1L, 2, universe = 1:3, bound = 3)

{1L [1], 2L [3], 3L [3]}
```

`gset_mean()` creates a new set by averaging corresponding memberships using the arithmetic, geometric or harmonic mean. Note that missing elements have 0 membership degree:

```
R> x <- gset(1:3, 1:3/3)
R> y <- gset(1:2, 1:2/2)
R> gset_mean(x, y)

{1L [0.4166667], 2L [0.8333333], 3L [0.5]}

R> gset_mean(x, y, "harmonic")

{1L [0.4], 2L [0.8]}

R> gset_mean(set(1), set(1, 2))

{1 [1], 2 [0.5]}
```

The membership vector of a generalized set can be transformed via `gset_transform_memberships()`, applying any *vectorized* function to the memberships:

```
R> x <- gset(1:10, 1:10/10)
R> gset_transform_memberships(x, pmax, 0.5)

{1L [0.5], 2L [0.5], 3L [0.5], 4L [0.5], 5L [0.5], 6L [0.6], 7L
 [0.7], 8L [0.8], 9L [0.9], 10L [1]}
```

Note that for multisets, an element's membership (multiplicity) m is interpreted as a one-vector of length m, yielding possibly unexpected results:

```
R> x <- gset(1, 2)
R> gset_transform_memberships(x, '*', 0.5)

{1 [{0.5 [2]}]}
```

For multisets, the `rep()` function is a more natural choice for membership transformations:

```
R> rep(x, 0.5)

{1}
```

In addition, three convenience functions are defined for fuzzy (multi-)sets: `gset_concentrate()` and `gset_dilate()` apply the square and the square root function, and `gset_normalize()` normalizes the memberships to a specified maximum:

```
R> gset_dilate(y)

{1L [0.7071068], 2L [1]}

R> gset_concentrate(y)

{1L [0.25], 2L [1]}

R> gset_normalize(y, 0.5)

{1L [0.25], 2L [0.5]}
```

(Note that these functions clearly have no effect for multisets.)

4.2. Fuzzy logic and fuzzy sets

For fuzzy (multi-)sets, the user can choose the logic underlying the operations using the `fuzzy_logic()` function. Fuzzy logics are represented as named lists with four components `N`, `T`, `S`, and `I` containing the corresponding functions for negation, conjunction ("t-norm"), disjunction ("t-conorm"), and (residual) implication (Klement, Mesiar, and Pap 2000). The fuzzy logic is selected by calling `fuzzy_logic()` with a character string specifying the fuzzy logic "family", and optional parameters. The exported functions `.N.()`, `.T.()`, `.S.()`, and `.I.()` reflect the currently selected bindings. Available families include: `"Zadeh"` (default), `"drastic"`, `"product"`, `"Lukasiewicz"`, `"Fodor"`, `"Frank"`, `"Hamacher"`, `"Schweizer-Sklar"`, `"Yager"`, `"Dombi"`, `"Aczel-Alsina"`, `"Sugeno-Weber"`, `"Dubois-Prade"`, and `"Yu"` (see Appendix A). A call to `fuzzy_logic()` without arguments returns the current logic.

```
R> x <- 1:10 / 10
R> y <- rev(x)
R> .S.(x, y)

 [1] 1.0 0.9 0.8 0.7 0.6 0.6 0.7 0.8 0.9 1.0

R> fuzzy_logic("Fodor")
R> .S.(x, y)

 [1] 1 1 1 1 1 1 1 1 1 1
```

Fuzzy set operations automatically use the active fuzzy logic setting:

```
R> X <- gset(c("A", "B", "C"), c(0.3, 0.5, 0.8))
R> print(X)

{"A" [0.3], "B" [0.5], "C" [0.8]}

R> Y <- gset(c("B", "C", "D"), c(0.1, 0.3, 0.9))
R> print(Y)

{"B" [0.1], "C" [0.3], "D" [0.9]}
```

First, we try the Zadeh logic (default):

```
R> fuzzy_logic("Zadeh")
R> X & Y

{"B" [0.1], "C" [0.3]}

R> X | Y

{"A" [0.3], "B" [0.5], "C" [0.8], "D" [0.9]}

R> gset_complement(X, Y)

{"B" [0.1], "C" [0.2]}
```

The results are different by switching to the Fodor logic:

```
R> fuzzy_logic("Fodor")
R> X & Y

{"C" [0.3]}

R> X | Y
```

```
{"A" [0.3], "B" [0.5], "C" [1], "D" [0.9]}
```

```
R> gset_complement(X, Y)
```

```
{}
```

The `cut()` method for generalized sets "filters" all elements with memberships exceeding a specified level (α-cuts)—the result, thus, is a crisp (multi)set:

```
R> cut(X, 0.5)
```

```
{"B", "C"}
```

The method can also be used for ν-cuts, selecting elements according to their multiplicity.

4.3. Characteristic functions and their visualization

The **sets** package provides several generators of characteristic functions to be used as templates for the creation of fuzzy sets, including the following shapes: gaussian curve (`fuzzy_normal()`), double gaussian curve (`fuzzy_two_normals()`), bell curve (`fuzzy_bell()`), sigmoid curve (`fuzzy_sigmoid()`), Π-like curves (`fuzzy_pi3()`, `fuzzy_pi4()`), trapezoid (`fuzzy_trapezoid()`), and triangle (`fuzzy_triangular()`, `fuzzy_cone()`). For example, a fuzzy normal function and a corresponding fuzzy set are created using:

```
R> N <- fuzzy_normal(mean = 0, sd = 1)
R> N(-3:3)
```

```
[1] 0.0111090 0.1353353 0.6065307 1.0000000 0.6065307 0.1353353 0.0111090
```

```
R> gset(charfun = N, universe = -3:3)
```

```
{-3L [0.011109], -2L [0.1353353], -1L [0.6065307], 0L [1], 1L
 [0.6065307], 2L [0.1353353], 3L [0.011109]}
```

For convenience, we also provide wrappers that directly generate corresponding sets, given a specified universe:

```
R> fuzzy_normal_gset(universe = -3:3)
```

```
{-3L [0.011109], -2L [0.1353353], -1L [0.6065307], 0L [1], 1L
 [0.6065307], 2L [0.1353353], 3L [0.011109]}
```

(If no universe is specified, the default universe is used; if this is also missing, the universe currently defaults to `seq(0, 20, 0.1)`.) It is also possible to create function generators for characteristic functions from other functions (such as distribution functions):

```
R> fuzzy_poisson <- charfun_generator(dpois)
R> gset(charfun = fuzzy_poisson(10), universe = seq(0, 20, 2))
```

```
{0 [0.00036288], 2 [0.018144], 4 [0.1512], 6 [0.504], 8 [0.9], 10
 [1], 12 [0.7575758], 14 [0.4162504], 16 [0.1734377], 18
 [0.05667898], 20 [0.01491552]}
```

`fuzzy_tuple()` generates a sequence (tuple) of sets based on any of the generating functions (except `fuzzy_trapezoid()` and `fuzzy_triangular()`). The chosen generating function is called with different values (chosen along the universe) passed to the first argument, thus varying the position or the resulting graph:

```
R> ## creating a series of fuzzy normal sets
R> fuzzy_tuple(fuzzy_normal, 5)

(<<gset(201)>>, <<gset(201)>>, <<gset(201)>>, <<gset(201)>>,
 <<gset(201)>>)
```

(`<<gset(201)>>` denotes an object of class '`gset`' with 201 elements—the size of the default universe). The **sets** package provides support for visualizing the membership functions of generalized sets, and in particular fuzzy sets. For (fuzzy) multisets, the plot method produces a (grouped) barplot for the membership vector (see Figure 1, top left):

```
R> ## a fuzzy multiset
R> X <- gset(c("A", "B"), list(1:2/2, 0.5))
R> plot(X)
```

Characteristic function generators can directly be plotted using a default universe (see Figure 1, top right):

```
R> plot(fuzzy_bell)
```

There is a plot method for tuples for visualizing a sequence of sets (see Figure 1, bottom left):

```
R> ## creating a sequence of sets
R> plot(fuzzy_tuple(fuzzy_cone, 10), col = gray.colors(10))
```

Plots of several sets can be superposed using the line method (see Figure 1, bottom right):

```
R> x <- fuzzy_normal_gset()
R> y <- fuzzy_trapezoid_gset(corners = c(5, 10, 15, 17), height = c(0.7, 1))
R> plot(tuple(x, y), lty = 3)
R> lines(x | y, col = 2)
R> lines(gset_mean(x, y), col = 3, lty = 2)
```

Finally, we note that the **sets** package provides basic infrastructure for fuzzy inference (`fuzzy_inference()`), involving the definition of linguistic variables (`fuzzy_variable()`, `fuzzy_partition()`) and fuzzy rules (`fuzzy_rule()`) to build a fuzzy system (`fuzzy_system()`). A few methods for defuzzification (`gset_defuzzify()`) are also provided.

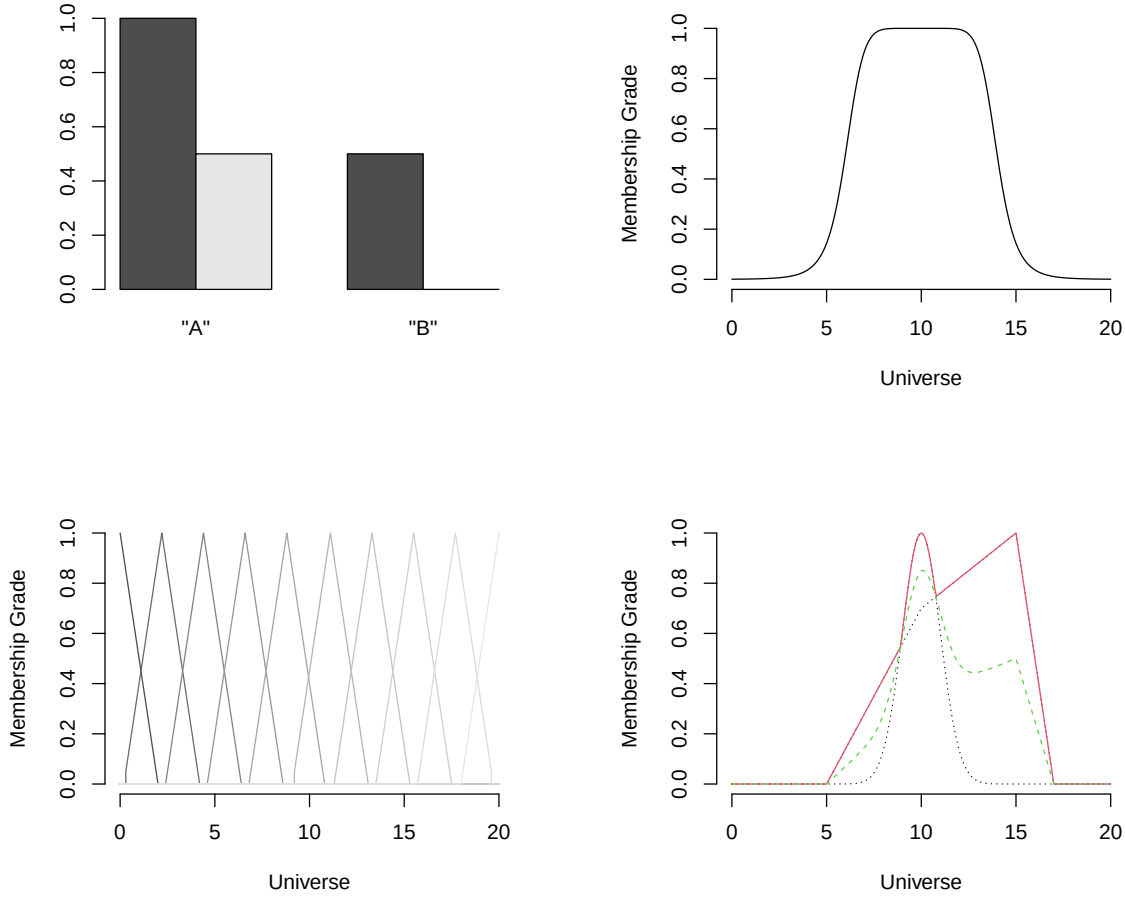

Figure 1: Membership plots for fuzzy sets. Top left: grouped barplot for a fuzzy multiset. Top right: graph of a bell curve. Bottom left: sequence of triangular functions. Bottom right: two combinations of a normal and a trapezoid function (dotted lines: basic shapes; solid (red) line: union; dashed (green) line: arithmetic mean).

5. User-definable extensions

We added *customizable sets* extending generalized sets in two ways: First, users can control the way elements are matched, i.e., define equivalence classes of elements. Second, arbitrary iteration orders can be specified.

5.1. Matching functions

By default, sets and generalized sets use `identical()` to match elements which is maximally restrictive. Note that this differs from the behavior of R's equality operator or `match()` which perform implicit type conversions and thus confound, e.g., `1`, `1L` and `"1"`. In the following example, note that on most computer systems, $3.3 - 2.2$ will not be identical to 1.1 due to

numerical issues.

```
R> x <- set("1", 1L, 1, 3.3 - 2.2, 1.1)
R> print(x)

{"1", 1L, 1, 1.1, 1.1}

R> y <- set(1, 1.1, 2L, "2")
R> print(y)

{"2", 2L, 1, 1.1}

R> 1L %e% y

[1] FALSE

R> x | y

{"1", "2", 1L, 2L, 1, 1.1, 1.1}
```

Customizable sets can be created using the `cset()` constructor, specifying the generalized set and some matching function.

```
R> X <- cset(x, matchfun = match)
R> print(X)

{"1", 1.1}

R> Y <- cset(y, matchfun = match)
R> print(Y)

{"2", 1, 1.1}

R> 1L %e% Y

[1] TRUE

R> X | Y

{"1", "2", 1.1}
```

Matching functions take two vector arguments, say, `x` and `table`, with `table` being a vector where the elements of `x` are looked up. The function should be vectorized in the `x`, i.e. return the first matching position for each element of `x`. In order to make use of non-vectorized predicates such as `all.equal()`, the **sets** package provides `matchfun()` to generate one:

```
R> FUN <- matchfun(function(x, y) isTRUE(all.equal(x, y)))
R> X <- cset(x, matchfun = FUN)
R> print(X)

{"1", 1L, 1.1}

R> Y <- cset(y, matchfun = FUN)
R> print(Y)

{"2", 2L, 1, 1.1}

R> 1L %e% Y

[1] TRUE

R> X | Y

{"1", "2", 1L, 2L, 1.1}
```

`sets_options()` can be used to conveniently switch the default match and/or order function if a number of '`cset`' objects need to be created:

```
R> sets_options("matchfun", match)
R> cset(x)

{"1", 1.1}

R> cset(y)

{"2", 1, 1.1}

R> cset(1:3) <= cset(c("1", "2", "3"))

[1] FALSE
```

5.2. Iterators

In addition to specifying matching functions, it is possible to change the order in which iterators such as `as.list()` (but not `for()`—see end of Section 3) process the elements. Note that the behavior of `as.list()` influences the labeling and print methods for customizable sets. Sets and generalized sets use the canonical internal ordering for iterations. With customizable sets, a "natural" ordering of elements can be kept by specifying either a permutation vector or an order function:

```
R> cset(letters[1:5], orderfun = 5:1)
```

```
{"e", "d", "c", "b", "a"}

R> FUN <- function(x) order(as.character(x), decreasing = TRUE)
R> Z <- cset(letters[1:5], orderfun = FUN)
R> print(Z)

{"e", "d", "c", "b", "a"}

R> as.character(Z)

[1] "e" "d" "c" "b" "a"
```

Note that converters for ordered factors keep the order:

```
R> o <- ordered(c("a", "b", "a"), levels = c("b", "a"))
R> as.set(o)

{a, b}

R> as.gset(o)

{a [2], b [1]}

R> as.cset(o)

{b [1], a [2]}
```

Converters for other data types will use the order information only if elements are unique:

```
R> as.cset(c("A", "quick", "brown", "fox"))

{"A", "quick", "brown", "fox"}

R> as.cset(c("A", "quick", "brown", "fox", "quick"))

{"A" [1], "brown" [1], "fox" [1], "quick" [2]}
```

6. Examples

In the following, we present two examples for the use of multisets and fuzzy multisets.

6.1. Multisets

Multisets are frequent in statistics since they can be seen as frequency tables of some objects. Using the **sets** package, a "generalized" table can easily be constructed from a list of R objects using the `as.gset()` coercion function. Assume, e.g., that one samples a number of fourfold tables given the margins using `r2dtable()`:

```
R> set.seed(4711)
R> l <- r2dtable(1000, r = 1:2, c = 2:1)
```

Since the sum of the first row (and second column) are constrained to 1, the top left cell entry can only be 0 or 1. Also, given the marginals, there is only one degree of freedom in fourfold tables, so the value of this first cell determines the others, and thus only two possible tables exist:

```
R> l[1:2]

[[1]]
     [,1] [,2]
[1,]    0    1
[2,]    2    0

[[2]]
     [,1] [,2]
[1,]    1    0
[2,]    1    1
```

To count them, we can simply use `as.gset()` that will construct a multiset from the list:

```
R> s <- as.gset(l)
R> print(s)

{<<2x2 matrix>> [330], <<2x2 matrix>> [670]}
```

Replace the matrices by the first cells' values:

```
R> for (i in s) s[[i]] <- i[1]
R> print(s)

{0L [330], 1L [670]}
```

The estimated probabilities of having 0 or 1 in the first cell can thus be obtained by:

```
R> gset_memberships(s) / 1000

   1    2
0.33 0.67
```

The probability for 0 clearly corresponds to the p value of the corresponding Fisher test:

```
R> fisher.test(l[[1]])$p.value

[1] 0.3333333
```

6.2. Fuzzy multisets

Fuzzy multisets can be used to represent objects appearing several times with different membership grades (e.g., weights, degrees of credibility, ...). Mizutani, Inokuchi, and Miyamoto (2008) describe an interesting application of fuzzy multisets to text mining: The occurrences of some terms of interest ("neural network", "fuzzy", "image") in titles, abstracts, and keywords of 30 documents on fuzzy theory are represented by fuzzy multisets, with varying memberships depending on whether a term occurs in the title (degree 1), the keywords (degree 0.6), and/or the abstract (degree 0.2).

```
R> data("fuzzy_docs")
R> print(fuzzy_docs[8:9])

$x8
{"fuzzy" [{0.2, 0.6}], "neural network" [{0.2, 1}]}

$x9
{"fuzzy" [{0.2, 0.6, 1}], "image" [{0.6}], "neural network" [{0.2,
 0.6, 1}]}
```

This information is then used to compute distances between documents, and ultimately to compare several (non-linear) clustering methods regarding their abilities of recovering the true underlying structure. In fact, it is known that the first 12 documents are related to neural networks, and the remaining 18 to image processing.

In the following, we will perform simple hierarchical clustering. We start by computing a distance matrix for the 30 documents. The **sets** package implements the Jaccard dissimilarity, defined for two generalized sets X and Y as $1 - |X \cap Y|/|X \cup Y|$ where $|\cdot|$ denotes the cardinality for generalized sets. A corresponding dissimilarity matrix can be obtained using, e.g., the **proxy** package (Meyer and Buchta 2009):

```
R> library("proxy")
R> d <- dist(fuzzy_docs, gset_dissimilarity)
```

We then apply Ward's clustering method:

```
R> cl1 <- hclust(d, "ward")
```

resulting in a clustering depicted in Figure 2. Clearly, the neural network-related documents (#1–#12) are separated from the image processing papers:

```
R> labs1 <- cutree(cl1, 2)
R> print(labs1)

 x1  x2  x3  x4  x5  x6  x7  x8  x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19
  1   1   1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2
x20 x21 x22 x23 x24 x25 x26 x27 x28 x29 x30
  2   2   2   2   2   2   2   2   2   2   2
```

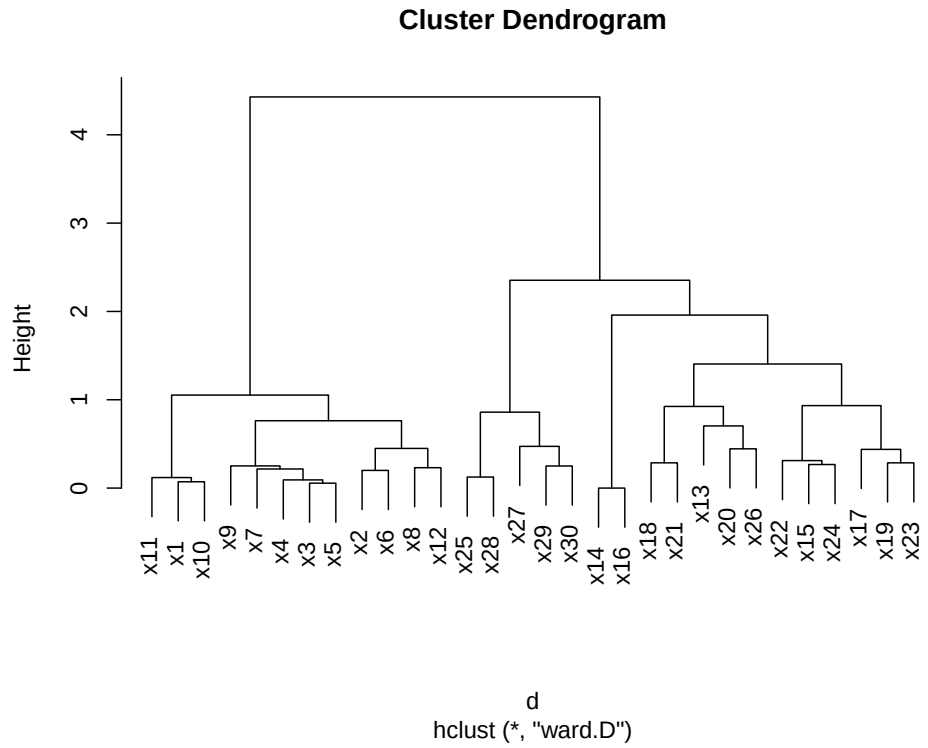

Figure 2: Dendrogram for the `fuzzy_docs` data, clustered using Ward's method on Jaccard distances computed from fuzzy multisets.

Note that for this data, using different weightings for terms in titles, keywords and abstracts are key to recover the subgroups. Naive text mining approaches operate on "classical" term-document-matrices only, counting term occurrences without further information on their relevance:

```
R> tdm <- set_outer(c("neural networks", "fuzzy", "image"),
+    fuzzy_docs, `%in%`)
R> print(tdm)

                   x1    x2    x3    x4    x5    x6    x7    x8    x9   x10
neural networks FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
fuzzy            TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
image           FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
                  x11   x12   x13   x14   x15   x16   x17   x18   x19   x20
neural networks FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
fuzzy            TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
image            TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
                  x21   x22   x23   x24   x25   x26   x27   x28   x29   x30
neural networks FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
fuzzy            TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
```

```
image            TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

By again computing Jaccard distances

```
R> d <- dist(tdm, "Jaccard", by_rows = FALSE)
```

and the corresponding hierarchical clustering, visualized in Figure 3,

```
R> cl2 <- hclust(d, "ward")
```

we can see that the "standard" text mining approach fails to correctly assign four documents (#9–#12):

```
R> labs2 <- cutree(cl2, 2)
R> print(labs2)

 x1  x2  x3  x4  x5  x6  x7  x8  x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19
  1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2
x20 x21 x22 x23 x24 x25 x26 x27 x28 x29 x30
  2   2   2   2   2   2   2   2   2   2   2

R> table(labs1, labs2)

     labs2
labs1  1  2
    1  8  4
    2  0 18
```

7. Conclusion

In this paper, we described the **sets** package for R, providing infrastructure for sets and generalizations thereof such as fuzzy sets, multisets and fuzzy multisets. The fuzzy variants make use of a dynamic fuzzy logic infrastructure offering several fuzzy logic families. Generalized sets are further extended to allow for user-defined iterators and matching functions. Current work focuses on data structures and algorithms for relations, an important application of sets.

References

Aklan S, Altindas E, Macit R, Umar S, Unal H (2008). ***fuzzyOP**: Fuzzy Numbers and the Main Mathematical Operations*. R package version 1.0, URL https://CRAN.R-project.org/package=fuzzyOP.

Alvarez AG (2007). ***FKBL**: Fuzzy Knowledge Base Learning*. R package version 0.50-4, URL https://CRAN.R-project.org/package=FKBL.

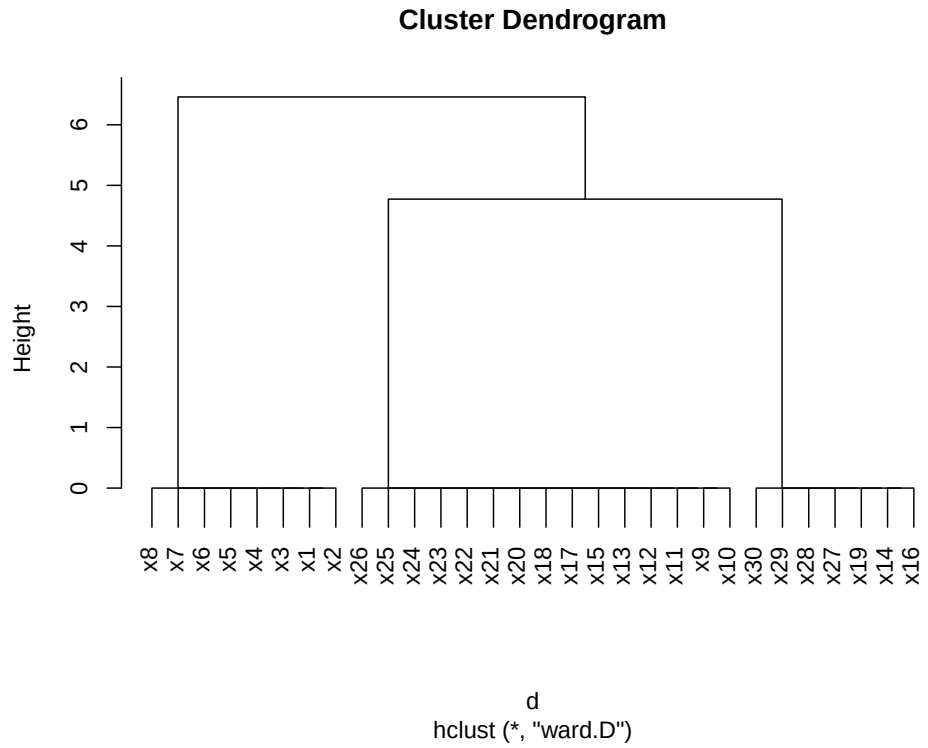

Figure 3: Dendrogram for the `fuzzy_docs` data, using Ward's method on a term-document-matrix generated for the data.

Cantor G (1895). "Beiträge zur Begründung der transfiniten Mengenlehre." In *Mathematische Annalen*, volume 46, pp. 481–512. Springer-Verlag.

Dubois D, Prade H, Yager RY (eds.) (1996). *Fuzzy Information Engineering: A Guided Tour of Applications*. John Wiley & Sons, New York.

Fraenkel AA (1922). "Über die Grundlagen der Cantor-Zermeloschen Mengenlehre." In *Mathematische Annalen*, volume 86, pp. 230–237. Springer-Verlag.

Geyer CJ (2007). ***fuzzyRankTests**: Fuzzy Rank Tests and Confidence Intervals*. R package version 0.3-2, URL https://CRAN.R-project.org/package=fuzzyRankTests.

Klement EP, Mesiar R, Pap E (2000). *Triangular Norms*. Springer-Verlag, Dordrecht.

Knuth DE (1973). *The Art of Computer Programming*, volume 3. Addison-Wesley, Reading.

Lewin A (2007). ***fuzzyFDR**: Exact Calculation of Fuzzy Decision Rules for Multiple Testing*. R package version 1.0, URL https://CRAN.R-project.org/package=fuzzyFDR.

Matthé T, Caluwe RD, de Tré G, Hallez A, Verstraete J, Leman M, Cornelis O, Moelants D, Gansemans J (2006). "Similarity Between Multi-valued Thesaurus Attributes: Theory and Application in Multimedia Systems." In *Flexible Query Answering Systems*, Lecture Notes in Computer Science, pp. 331–342. Springer-Verlag.

Meyer D, Buchta C (2009). ***proxy**: Distance and Similarity Measures*. R package version 0.4-3, URL https://CRAN.R-project.org/package=proxy.

Meyer D, Hornik K (2009a). "Generalized and Customizable Sets in R." *Journal of Statistical Software*, **31**(2), 1–27. doi:10.18637/jss.v031.i02.

Meyer D, Hornik K (2009b). ***sets**: Sets, Generalized Sets, and Customizable Sets*. R package version 1.0, URL https://CRAN.R-project.org/package=sets.

Mizutani K, Inokuchi R, Miyamoto S (2008). "Algorithms of Nonlinear Document Clustering Based on Fuzzy Multiset Model." *International Journal of Intelligent Systems*, **23**, 176–198.

Roberts DW (2007). ***fso**: Fuzzy Set Ordination*. R package version 1.0-1, URL https://CRAN.R-project.org/package=fso.

Singh D, Ibrahim A, Yohanna T, Singh J (2007). "An Overview of the Applications of Multisets." *Novi Sad Journal of Mathematics*, **37**(3), 73–92.

Vinterbo SA (2007). ***gcl**: Compute a Fuzzy Rules or Tree Classifier from Data*. R package version 1.06.5, URL https://CRAN.R-project.org/package=gcl.

Wirth N (1983). *Algorithmen und Datenstrukturen*. Teubner, Stuttgart.

Yager RR (1986). "On the Theory of Bags." *International Journal of General Systems*, **13**, 23–37.

Zadeh LA (1965). "Fuzzy Sets." *Information and Control*, **8**(3), 338–353.

Zermelo E (1908). "Untersuchungen über die Grundlagen der Mengenlehre." In *Mathematische Annalen*, volume 65, pp. 261–281. Springer-Verlag.

A. Available fuzzy logic families

Let us refer to $N(x) = 1 - x$ as the *standard* negation, and, for a t-norm T, let $S(x, y) = 1 - T(1 - x, 1 - y)$ be the *dual* (or complementary) t-conorm. Available specifications and corresponding families are as follows, with the standard negation used unless stated otherwise.

`"Zadeh"` Zadeh's logic with $T = \min$ and $S = \max$. Note that the minimum t-norm, also known as the Gödel t-norm, is the pointwise largest t-norm, and that the maximum t-conorm is the smallest t-conorm.

`"drastic"` The drastic logic with t-norm $T(x, y) = y$ if $x = 1$, x if $y = 1$, and 0 otherwise, and complementary t-conorm $S(x, y) = y$ if $x = 0$, x if $y = 0$, and 1 otherwise. Note that the drastic t-norm and t-conorm are the smallest t-norm and largest t-conorm, respectively.

`"product"` The family with the product t-norm $T(x, y) = xy$ and dual t-conorm $S(x, y) = x + y - xy$.

`"Lukasiewicz"` The Łukasiewicz logic with t-norm $T(x, y) = \max(0, x + y - 1)$ and dual t-conorm $S(x, y) = \min(x + y, 1)$.

`"Fodor"` The family with Fodor's *nilpotent minimum* t-norm given by $T(x,y) = \min(x,y)$ if $x + y > 1$, and 0 otherwise, and the dual t-conorm given by $S(x,y) = \max(x,y)$ if $x + y < 1$, and 1 otherwise.

`"Frank"` The family of Frank t-norms T_p, $p \geq 0$, which gives the Zadeh, product and Łukasiewicz t-norms for $p = 0$, 1, and ∞, respectively, and otherwise is given by $T(x,y) = \log_p(1 + (p^x - 1)(p^y - 1)/(p - 1))$.

`"Hamacher"` The three-parameter family of Hamacher, with negation $N_\gamma(x) = (1 - x)/(1 + \gamma x)$, t-norm $T_\alpha(x,y) = xy/(\alpha + (1 - \alpha)(x + y - xy))$, and t-conorm $S_\beta(x,y) = (x + y + (\beta - 1)xy)/(1 + \beta xy)$, where $\alpha \geq 0$ and $\beta, \gamma \geq -1$. This gives a deMorgan triple (for which $N(S(x,y)) = T(N(x), N(y))$ iff $\alpha = (1 + \beta)/(1 + \gamma)$.

The following parametric families are obtained by combining the corresponding families of t-norms with the standard negation and complementary t-conorm.

`"Schweizer-Sklar"` The Schweizer-Sklar family T_p, $-\infty \leq p \leq \infty$, which gives the Zadeh (minimum), product and drastic t-norms for $p = -\infty$, 0, and ∞, respectively, and otherwise is given by $T_p(x,y) = \max(0, (x^p + y^p - 1)^{1/p})$.

`"Yager"` The Yager family T_p, $p \geq 0$, which gives the drastic and minimum t-norms for $p = 0$ and ∞, respectively, and otherwise is given by $T_p(x,y) = \max(0, 1 - ((1 - x)^p + (1 - y)^p)^{1/p})$.

`"Dombi"` The Dombi family T_p, $p \geq 0$, which gives the drastic and minimum t-norms for $p = 0$ and ∞, respectively, and otherwise is given by $T_p(x,y) = 0$ if $x = 0$ or $y = 0$, and $T_p(x,y) = 1/(1 + ((1/x - 1)^p + (1/y - 1)^p)^{1/p})$ if both $x > 0$ and $y > 0$.

`"Aczel-Alsina"` The family of t-norms T_p, $p \geq 0$, introduced by Aczél and Alsina, which gives the drastic and minimum t-norms for $p = 0$ and ∞, respectively, and otherwise is given by $T_p(x,y) = \exp(-(|\log(x)|^p + |\log(y)|^p)^{1/p})$.

`"Sugeno-Weber"` The family of t-norms T_p, $-1 \leq p \leq \infty$, introduced by Weber with dual t-conorms introduced by Sugeno, which gives the drastic and product t-norms for $p = -1$ and ∞, respectively, and otherwise is given by $T_p(x,y) = \max(0, (x+y-1+pxy)/(1+p))$.

`"Dubois-Prade"` The family of t-norms T_p, $0 \leq p \leq 1$, introduced by Dubois and Prade, which gives the minimum and product t-norms for $p = 0$ and 1, respectively, and otherwise is given by $T_p(x,y) = xy/\max(x,y,p)$.

`"Yu"` The family of t-norms T_p, $p \geq -1$, introduced by Yu, which gives the product and drastic t-norms for $p = -1$ and ∞, respectively, and otherwise is given by $T(x,y) = \max(0, (1 + p)(x + y - 1) - pxy)$.

Affiliation:

David Meyer
Department of Information Systems and Operations
WU Wirtschaftsuniversität Wien
Augasse 2–6
1090 Wien, Austria E-mail: David.Meyer@wu.ac.at
URL: http://wi.wu.ac.at/~meyer/

Kurt Hornik
Department of Statistics and Mathematics
WU Wirtschaftsuniversität Wien
Augasse 2–6
1090 Wien, Austria E-mail: Kurt.Hornik@wu.ac.at
URL: http://statmath.wu.ac.at/~hornik/