Multisets

Luciano da Fontoura Costa

Multisets

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

15th Oct 2021

Abstract

Multisets are sets that allow repetition of elements, therefore accounting for their frequency, or multiplicity of observation. As such, multisets provide flexible resources for scientific modeling. In the present work, after revising the main aspects of traditional sets, we introduce some of the main concepts and characteristics of multisets, which is followed by their generalization to take into account vectors and matrices, and then functions and scalar and vector fields. These developments require multisets to become capable of coping with negative multiplicities, which gives rise to several additional set operations. Then multiset operations can be naturally incorporated into real function spaces allowing, among other possibilities, the definition of a De Morgan theorem between real-valued functions. Special attention is given to understanding the Jaccard and coincidence similarity indices in the context of real-valued multisets and functions, and it is shown that these indices, especially the latter, can yield narrow and sharp peaks corresponding to pattern matchings while attenuating secondary structures.

'In the bag, seashells gathered long ago resound.'

LdaFC

1 Introduction

Multisets (e.g. [1, 2, 3, 4, 5, 6]) — henceforth *msets* — can be informally understood as sets allowing repeated entries of the same element. In a sense, they are at least as much compatible with human intuition than sets. For instance it is often more relevant to know that our bag contains 4 apples than knowing simply that there are only apples in it. Given their enhanced potential for representing real-world structures and dynamics, msets result provide particularly effective resources for scientific modeling (e.g. [7]).

The present work has three main objectives: (i) to present an introduction to msets; (ii) develop extensions and generalizations to real-valued functions and scalar and vector fields; and (iii) illustrate the potential of the presented and proposed concepts and methods with respect to applications of the real-valued Jaccard and coincidence indices. The latter index [8] integrates the information provided by the Jaccard and interiority indices, therefore providing a more strict and detailed quantification of the similarity between the two compared msetes.

Special attention is given to extending element multiplicity to negative values, which is necessary condition for generalizing msets to functions and fields. This generalization has several interesting effects, such as allowing the important complement operation to be performed by using the empty mset instead of the universe mset. In addition, several additional mset operations can be defined respectively to how the positive and negative multiplicities are to be taken into account, some of which directly related to the Jaccard and coincidence similarity indices [8].

We start by reviewing the main concepts and properties of traditional sets, and then present the concept of msets, as well as some of their simpler properties, also including several examples. The challenges implied by the definition of a universe set for musets is briefly characterized and discussed. It is also argued that the operations of sum and subtraction between msets correspond to one of the main distinction between multiset and set theories.

The possibility to generalize msets to several other mathematical structures including vectors, matrices, functions, scalar and vector fields, as well as probability densities are approached next, including several examples. In particular, the extension of msets as representations of functions and scalar fields paves the way for obtaining hybrid expressions involving combinations of the the set operations of union and intersection with algebraic expressions involving sum, subtraction, product and division of sets. We illustrate the potential of this approach

by extending the De Morgan theorem to functions and fileds.

The interesting possibility to approach similarity indices such as the Jaccard and coincidence in terms of msets is then developed. We then presents how the definition of the common product between two msets or mfunctions paves the way to obtaining integrated signal operations including filtering and enhanced template matching. In particular, we illustrate the enhanced potential of the real-valued Jaccard and coincidence indices for template matching, an operation frequently performed in the areas of pattern recognition and neuronal networks. When compared to the classic cross-correlation, the real-valued Jaccard, and particular the coincidence index are verified to yield substantially sharper and narrower peaks indicating the position of the pattern matches, while secondary smaller scale structures are attenuated. This important property paves the way to several related applications of the coincidence index in artificial intelligence, deep learning, and scientific modeling.

The application of msets and the Jaccard index to quantify the relationship between two or more densities or clusters is then described with respect to an example related to the iris dataset. The measurement of the separation between clusters corresponds to an important issue in both pattern recognition (e.g. [9, 10]), deep learning (e.g. [11, 12, 13]), and modeling (e.g. [14, 7])

For simplicity's sake, the term msets are henceforth abbreviated as *msets*.

2 Traditional Sets

A *set* (e.g. [1, 15]) is an unordered collection of items, or *elements*, which are *not* allowed to repeat. A set A with elements a, b, and c is typically represented as:

$$A = \{a, b, c\}$$

The two essential properties of sets therefore are that the elements may appear in any order, which distinguish sets from vectors, and that the elements cannot be repeated.

Observe that a set can also have sets as elements. The number of elements in a set is called its *cardinality* or *size*, being represented as $|A|$.

A *subset* B of a given set A consists of a set so that any of its elements belong to A. If A contains N elements, there will be $|A| = 2^N$ possible subsets that can be derived from it. The set containing all possible subsets of A is called its *power set* P^A.

An important point about sets that is sometimes overlooked regards the fact that they always refer to a respective *universe set* Ω. More specifically, once this set is established, any possible set needs to be a subset of Ω. Observe that Ω can have any type of elements, though the situation where the elements are homogenous (e.g. positive integers, or real values) is of particular interest.

In case some sets are given but the universe set is not provided, it is still possible to estimate the respective universe set as corresponding to the union of all the existing sets.

The universe set is of fundamental importance because the operation of *complement* of a set is defined with respect to it. More specifically, the complement of a set A consists of all elements of Ω that are not in A. The complement of a set is henceforth represented as A^C, being implicit that the operation refers to a given Ω.

Sets can be finite or infinite, as well as discrete or continuous. A *finite set* is any set A so that $|A| < \infty$. A *discrete set* is characterized by having all its elements corresponding to isolated points p. Any continuous set is infinite, but discrete sets can be finite or infinite.

An interesting point regards the relationship between an element, let's say 'a' and the set $\{a\}$. These two mathematical structures are not identical because it is possible to include an element into $\{a\}$, but not into 'a'.

The *empty set*, represented as $\phi = \{\}$ is a subset of any possible set.

Given two sets A and B, their *union* consists of a third set C containing all elements from A and B. The *intersection* of these two sets corresponds to a set C containing all elements that are in both A and B. A subset B of A can therefore be understood as to be so that $A \cap B = A$. Any set is a subset of itself.

The *difference* between two sets A and B, indicated as $A - B$, corresponds to the set C containing all elements that are in A but are not in B.

Given three sets A, B, and C derived from a given Ω, the following properties directly involving the universe and empty set are verified:

$$\Omega^C = \Omega - \Omega = \Phi \tag{1}$$
$$\Phi^C = \Omega - \Phi = \Omega \tag{2}$$
$$A^C = \Omega - A \tag{3}$$
$$\Omega \cup \Phi = \Omega \tag{4}$$
$$\Omega \cap \Phi = \Phi \tag{5}$$
$$A \cup A^C = \Omega \tag{6}$$
$$A \cap A^C = A \tag{7}$$
$$A \cup \Phi = A \tag{8}$$
$$A \cap \Phi = \phi \tag{9}$$

Additional operations involving one or two sets A and

B include the following:

$$A \cup A = A \tag{10}$$
$$A \cap A = A \tag{11}$$
$$A \cup B = B \cup A \tag{12}$$
$$A \cap B = B \cap A \tag{13}$$
$$A \cup B^C = \Omega - (B - A) \tag{14}$$
$$A \cap B^C = A - B \tag{15}$$

And operations involving three sets A, B and C include:

$$A \cup (B \cup C) = (A \cup B) \cup C \tag{16}$$
$$A \cap (B \cup C) = (A \cap B) \cup C \tag{17}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{18}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{19}$$
$$(A \cup B)^C = A^C \cap B^C \text{ (De Morgan)} \tag{20}$$
$$(A \cap B)^C = A^C \cup B^C \text{ (De Morgan)} \tag{21}$$

Examples of continuous sets include intervals along the real line, e.g. $[0, 1]$ and $(b, c]$, and regions in $\mathbb{R}^2$, such as (x, y) satisfying $\sqrt{x^2 + y^2} \leq r$.

3 Multisets

Basically, *msets* are sets allowing the *repetition* of elements, which is understood as their *multiplicity* or *frequency*. As with sets, the order of the elements is immaterial. Examples of msets include:

$$\begin{aligned} A &= \{\!\{a, a, b, b, b, d\}\!\}; \\ B = \{\!\{1, 2, 1, 2, 1, 2, 1\}\!\} &= \{\!\{1, 1, 1, 1, 2, 2, 2\}\!\}; \\ C = \{\!\{1, a, 2, b, b, 3, c, c, c, 1, d, 2, a, a\}\!\} &= \\ = \{\!\{1, 1, 2, 2, 3, a, a, a, b, b, c, c, c\}\!\}; \\ D &= \{\!\{a, a, b, d\}\!\}. \end{aligned}$$

Observe the different symbol adopted henceforth in this work in order to emphasize the distinction between a traditional set ($\{\}$), and a mset ($\{\!\{\}\!\}$).

A more compact representation of a mset A can be obtained by using 2-tuple or pairs $[a, m(a)]$, where 'a' is an element and $m(a)$ it its multiplicity, i.e. the number of times it appear in A. In the case of the above examples, we have:

$$\begin{aligned} A = \{\!\{a, a, b, b, b, d\}\!\} &= \{\!\{[a, 2]; [b, 2]; [d, 1]\}\!\}; \\ B = \{\!\{1, 1, 1, 1, 2, 2, 2\}\!\} &= \{\!\{[1, 4]; [2, 3]\}\!\}; \\ C = \{\!\{1, 1, 2, 2, 3, a, a, a, b, b, c, c, c\}\!\} &= \\ = \{\!\{[1, 2]; [2, 2]; [3, 1]; [a, 3]; [b, 2]; [c, 3]\}\!\}; \\ D = \{\!\{a, a, b, d\}\!\} &= \{\!\{[a, 2]; [b, 1]; [d, 1]\}\!\}. \end{aligned} \tag{22}$$

Though this type of representation of msets actually corresponds to a set, because it is impossible to have two identical entries, we shall maintain the '$\{\!\{\}\!\}$' notation in order to emphasize that a mset is being meant.

When referring to the multiplicity of an element, it is important to specify to which mset this is being referred. This can be done by writing $m_A(a)$, meaning the multiplicity of the element a in the mset A.

The property analogous to inclusion in sets can be stated as follows. A mset A is included in another mset B whenever $m_A(a) \leq m_B(a)$. For instance, in the case of the examples above, we have $m_A(a) = 2$ and $m_C(a) = 2$.

The *support* of a given mset A is defined as:

$$S_A = \{x | x \in \Omega, m(x) > 0\} \tag{23}$$

As such, this set can be understood as containing all distinct elements in A. Observe that the support set provides a useful index for identifying the possible elements in the respective msets.

For instance, the supports of the msets in Equation 22

$$\begin{aligned} S_A &= \{a, b\} \\ S_B &= \{1, 2\} \\ S_C &= \{1, 2, a, b, c\} \\ S_D &= \{a, b, d\} \end{aligned} \tag{24}$$

The combined support of two msets is the union of their respective supports. Thus, in the case of the previous example, we have:

$$S_{A,B} = \{a, b, d, 1, 2\} \tag{25}$$
$$S_{A,B,C,D} = \{a, b, c, d, 1, 2\} \tag{26}$$

4 Multiset Operations

A set A is said to be *included* into another set whenever:

$$m_A(x) \leq m_b(x), \forall x \in A \tag{27}$$

For simplicity's sake, we will indicate this operation using the same symbol as for sets, i.e. $A \subseteq B$, as the type of operation can be inferred from A and B being sets or msets.

In the case of the mset examples above, we can write that $D \subseteq A$.

The *union* C of two msets A and B can be defined as:

$$\begin{aligned} C = A \cup B = \{\!\{[x, m_C(x)], x \in S_{A,B}\}\!\}, \\ \text{with } m_C(x) = \max\{m_A(x), m_B(x)\} \end{aligned} \tag{28}$$

Examples considering the msets in the beginning of Section 3 include:

$$\begin{aligned} A \cup B = \{\!\{a, a, b, b, b, d, 1, 1, 1, 1, 2, 2, 2\}\!\} &= \\ \{\!\{[1, 2]; [b, 2]; d, 1]; 1, 4]; [2, 3]\}\!\} \\ A \cup D = \{\!\{a, a, b, d\}\!\} &= \{\!\{[a, 4]; [b, 4]; [d, 2]\}\!\} \end{aligned}$$

Let A and B be msets. The *sum* of these two sets, henceforth represented as $C = A + B$, is defined as:

$$C = A + B = \{\!\{[x, m_C(x)], x \in S_{A,B}\}\!\}, \\ \text{with } m_C(x) = m_A(x) + m_B(x) \tag{29}$$

Figure 1 illustrates the two different ways in which the common elements of two msets A and B are collected into their respective union and sum msets.

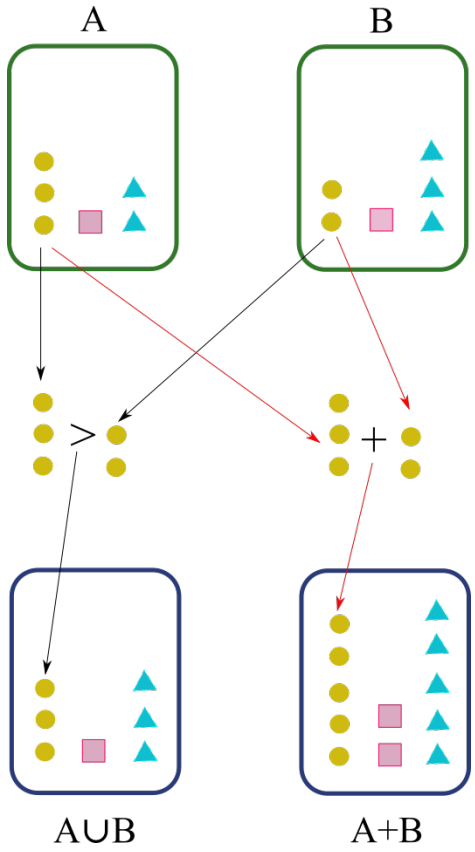


Figure 1: The union (a) and sum (b) of two msets A and B typically yield different resulting msets. In the case of the union operation, each of the elements of the same type are compared, with the elements with the maximum multiplicity being incorporated into C. The sum of the two msets incorporates all the $m_A(x_i) + m_B(x_i)$ elements into C.

Examples respective to the msets in the beginning of Section 3 include:

$$A + B = \{\!\{a, a, b, b, b, d, 1, 1, 1, 1, 2, 2, 2\}\!\} = \\ \{\!\{[1, 2]; [b, 2]; [d, 1]; 1, 4]; [2, 3]\}\!\} \\ A + D = \{\!\{a, a, a, a, b, b, b, b, d, d\}\!\} = \{\!\{[a, 4]; [b, 4]; [d, 2]\}\!\}$$

Thus, we have that the mset operations of union and sum are related in the sense that both collect the elements from the two msets, but the way in which this is done is quite different, with the multiplicities of the mset obtained by union becoming necessarily smaller or equal than that of the mset obtained by sum, i.e. $m_{A \cup B}(x_i) \leq m_{A+B}$.

It is interesting to consider these two operations in the context of possible respective applications. The sum of the two msets ensures conservation of the total number of elements (such as in conservative or flow-related problems), being therefore more indicated for related situations. The union of two msets can be conceptually understood as a choice procedure which does not ensure conservation of the multiplicities.

The *intersection* between two msets A and B can be defined as:

$$C = A \cap B = \{\!\{[x, m_C(x)], x \in S_{A,B}\}\!\}, \\ \text{with } m_C(x) = \min\{m_A(x), m_B(x)\} \tag{30}$$

Examples drawn from the msets in the beginning of Section 3 include:

$$A \cap B = \{\!\{\}\!\} \\ A \cap D = \{\!\{a, a, b\}\!\} = \{\!\{[a, 4]; [b, 4]; [d, 2]\}\!\} \tag{31}$$

The *difference* or *subtraction* between two msets is expressed as:

$$C = A - B = \{\!\{[x, m_C(x)], x \in S_{A,B}\}\!\}, \\ \text{with } m_C(x) = \max\{m_A(x) - m_B(x), 0\} \tag{32}$$

In Section 8, we will show that allowing negative multiplicities paves the way to defining the mset complement as well as to several additional mset operations.

Figure 2 illustrates the intersection and difference between two msets A and B.

Respective examples include:

$$A - D = \{\!\{b, b\}\!\} = \{\!\{[b, 2]\}\!\} \\ D - A = \{\!\{\}\!\} \tag{33}$$

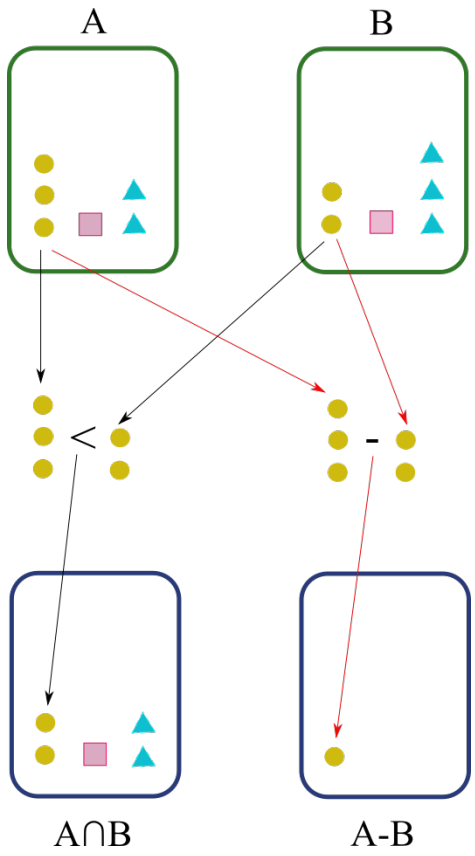


Figure 2: The intersection (a) and difference (b) of two msets A and B typically yield quite different resulting msets. In the case of the intersection operation, each of the elements of the same type are compared, with the elements with the minimum respective multiplicity being incorporated into C. The difference between A and B depends on $m_A(x_i) - m_B(x_i)$. As the result is negative in the case of the present example, no elements are incorporated into $A - B$.

5 Multisets Properties

It can be shown that msets as presented in the previous section satisfy the following properties:

$$A \cup \Phi = A \tag{34}$$
$$A \cap \Phi = \Phi \tag{35}$$
$$A \cup A = A \tag{36}$$
$$A \cap A = A \tag{37}$$
$$A \cup B = B \cup A \tag{38}$$
$$A \cap B = B \cap A \tag{39}$$
$$A \cup (B \cup C) = (A \cup B) \cup C \tag{40}$$
$$A \cap (B \cup C) = (A \cap B) \cup C \tag{41}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{42}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{43}$$

where Φ is the *empty* mset, which contains null multiplicities for all elements in the respective support. This set can be expressed as:

$$\Phi = \{[x_i, 0]\}\, i = 1, 2, \ldots, |S| \tag{44}$$

So, we have that msets follow all the properties in Equations 8 to Equations 21 , except those involving complementation.

The definition of the complement of an mset has been a challenging issue (e.g. [6]), which has to do with the fact that, typically, multiplicities have been restricted to non-negative integer multiplicities (see also [16] and [17]). In particular, restricting the mset difference operation multiplicities to take only non-negative values makes it difficult to define a respective complement operation. For this reason, several useful De Morgan properties, as well as other related results, are not extended msets.

6 The Multiset Jaccard Indices

The Jaccard index represents an effective and conceptually appealing manner to quantify the similarity between any two sets A and B (e.g. [18, 19, 20, 21, 22, 23, 24, 25, 26, 23, 27, 28, 29, 8]), having therefore being extensively applied in a vast range of problems in several scientific and technological fields.

In its most basic form, the Jaccard index between an two sets A and B can be expressed as:

$$\mathcal{J}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{45}$$

It is possible to adapt the Jaccard index to msets by making:

$$\mathcal{J}_M(A, B) = \frac{\sum_{i=1}^{N} \min\left(m(a_i), m(b_i)\right)}{\sum_{i=1}^{N} \max\left(m(a_i), m(b_i)\right)} \tag{46}$$

where a_i and b_i are the elements of the sets A and B, respectively, and N is the cardinality of the universe of those sets. We also have that $0 \leq \mathcal{J}_M(A, B) \leq 1$.

As an example, let's consider $A = \{a, b, b, c, c, c\}$ and $B = \{a, b, c, c, d\}$. Then, we have:

$$\mathcal{J}(A, B) = \frac{1 + 1 + 2 + 0}{2 + 3 + 5 + 1} = \frac{4}{11} \tag{47}$$

It is possible to adapt the Jaccard index to mfunctions by making:

$$\mathcal{J}(A, B) = \frac{\int_{\Phi} \min\left(m_A(\vec{x}), m_B(\vec{x})\right)}{\int_{\Phi} \max\left(m_A(\vec{x}), m_B(\vec{x})\right)} \tag{48}$$

where Φ is the common support of the two functions or scalar fields, and $0 \leq \mathcal{J}(A, B) \leq 1$.

As such, the Jaccard index can be understood as a *functional*, or *mfunctional* of the two functions of scalar fields.

The Jaccard index has been enhanced and extended to functions, scalar fields, joint variations and more than 2 sets [8]. In particular, the latter type of Jaccard index for 3 sets A, ,B and C can be written as:

$$\mathcal{J}(A,B,C) = \frac{A \cap B \cap C}{A \cup B \cup C} \tag{49}$$

7 Multisets, Vectors, Matrices

In this section we will discuss the relationship between msets and vectors. Observe that a vector can be understood as an mset with real multiplicities.

First, we recall that the elements in a vector are expected to follow a well-determined order as indicated by their indices. For instance, in the case of the vector in $\mathbb{R}^5$:

$$\vec{v} = [3, 2.5, \pi, 0, -1]$$

we have five indices $i = 1, 2, \ldots, 5$, so that we can specify the respective element values as $v[1] = 3$, $v[2] = 2.5$, $v[3] = \pi$, $v[4] = 0$, and $v[5] = -1$.

By understanding the values of the components of a vector as generalized multiplicities, we can immediately derive the following mset from the above vector:

$$V = \{\!|[1,3]; [2,2.5]; [3,\pi]; [4,0]; [5,-1]|\!\}$$

More generally, we have that a vector $\vec{v}$ can be bijectively represented by the following mset:

$$\vec{v} = \{[i, v[i]]\} \tag{50}$$

Therefore, we have that an mset can be derived from any vector, but that a vector can be obtained from an mset only if their elements are ordered in some manner, e.g. by taking their respective values instead of understanding them as labels. This situation becomes more evident when one considers non-numeric elements. As such, msets can be used to study the elements of vectors without taking into account their relative position along the vector.

It is also interesting to contemplate the relationship between the above discussion and the traditional sets containing multiplicities. More specifically, we can write the set containing all multiplicities in the vector $\vec{v}$ above as:

$$\tilde{V} = \{3, 2.5, \pi, 0, -1\} = \{0, 3, 2.5, -1, \pi\} = \text{etc.}$$

Though V and $\tilde{V}$ are very similar, they are not identical because in V the *correspondence* of the elements and the respective multiplicity is maintained by the specification of the respective element in the ordered pair. By representing vectors as msets, we not only preserve the operations of subtraction and difference, but also incorporates the possibilities of defining intersections and unions between any two vectors.

Another interesting possibility is to incorporate new operators for multiplication and division into the mset framework, which can be done straightforwardly, while avoiding divisions by zero.

Interestingly, it is also possible to obtain mset representations from matrices or even other more sophisticated mathematical structures as tensors. In the case of matrices, this can be done by incorporating the indexing information in the respective mset. For instance, an $N \times M$ matrix can be represented as the mset:

$$A = \{[i, j, A[i,j])]\} \tag{51}$$

An additional point remains to be discussed regarding the fact that vectors, matrices and other mathematical structures are restricted to non-negative integer entries.

In the present work, we propose the mset difference operation to be modified as:

$$\begin{aligned} C = A - B = \{\!|[x, m_C(x)], x \in S_{A,B}|\!\}, \\ \text{with } m_C(x) = m_A(x) - m_B(x) \end{aligned} \tag{52}$$

It is now possible to extend the multiplicities to take any real value.

While the empty is as observed before, i.e. corresponding to an mset with null multiplicities for every element in the support, we now need to identify a suitable universe set. In the case of vectors with dimension $1 \times N$ this can be done by making:

$$\Omega_+ = \{[i, \infty]\}, i = 1, 2, \ldots N \tag{53}$$

$$\Omega_- = \{[i, -\infty]\}, i = 1, 2, \ldots N \tag{54}$$

Now, we have that the complement of an mset can be expressed as:

$$A^C = \Phi - A \tag{55}$$

Observed that the empty mset has been used instead of the universe mset.

We can now incorporate the following additional properties to real-valued msets:

$$\Omega_+^C = \Omega_+ - \Omega_+ = \Phi \tag{56}$$

$$\Omega_-^C = \Omega_- - \Omega_+ - = \Phi \tag{57}$$

$$A^C = \Phi - A = -A \tag{58}$$

$$\Omega_+ \cup \Phi = \Omega_+ \tag{59}$$

$$\Omega_- \cup \Phi = \Phi \tag{60}$$

$$\Omega_+ \cap \Phi = \Phi \tag{61}$$

$$\Omega_- \cap \Phi = \Omega_- \tag{62}$$

$$A \cup A^C = |A| \tag{63}$$

$$A \cap A^C = -|A| \tag{64}$$

We can also define the following operations on a generic mset A:

$$A_+ = \{\!|[x, m_A(x)], x \in S_A|\!\},$$
$$\text{with } m_A(x) = \max\{m_A(x), 0\} \tag{65}$$

and

$$A_- = \{\!|[x, m_A(x)], x \in S_A|\!\},$$
$$\text{with } m_A(x) = \min\{m_A(x), 0\} \tag{66}$$

which allow us to write:

$$A \cup \Phi = A_+ \tag{67}$$
$$A \cap \Phi = A_- \tag{68}$$

The fact that $A \cup \Phi$ is not A and $A \cap \Phi$ is not Φ as in traditional set theory motivates us to find a new mset operation for real-valued msets that could lead to a respective counterpart. This can be done by defining the *signed union* of two msets A and B as follows:

$$C = A \sqcup B =$$
$$= \{\!|[x, s_{m_A(x)} s_{m_B(x)} \max\{|m_A(x)|, |m_B(x)|\}]|\!\}$$
$$\text{with } x \in S_{A,B} \tag{69}$$

where $s_{m_A(x)}$ stands for the sign of the multiplicity $m_A(x)$ and $s_{m_B(x)}$ stands for the sign of the multiplicity $m_B(x)$.

The *signed intersection* can be expressed as:

$$C = A \sqcap B =$$
$$= \{\!|[x, s_{m_A(x)} s_{m_B(x)} \min\{|m_A(x)|, |m_B(x)|\}]|\!\}$$
$$\text{with } x \in S_{A,B} \tag{70}$$

The above expression was previously reported [30] in the context of cosine similarity analogous to the L1 norm.

Now, we can write:

$$A \sqcup \Phi = A \tag{71}$$
$$A \sqcap \Phi = \Phi \tag{72}$$

Observe that the consideration of possibly negative multiplicities has led to additional respective properties.

8 Functions and Scalar Fields

The possibility to represent vectors as msets paves the way to a number of interesting possibilities. One of them is to represent discrete and continuous *functions* and *scalar fields* (vector fields can be approached as vectors of scalar fields). We develop these possibilities in the following.

Let $g(x)$ be a real function of a real variable x in a discrete or continuous support set S, i.e.$x \in S$. This function can be fully represented, in invertible manner, in terms of the following respective real-valued mset:

$$f(x) = f = \{[x, f(x)]\}, \text{ with } x \in S \tag{73}$$

Therefore, the extension of the above characterized real-valued msets to real-valued function is immediate. Though for simplicity's sake we shall often call the msets associated from functions as *mfunctions*, it should be kept in mind that there is absolutely no difference between real-valued msets and real functions.

All the respective operations and properties are also kept, which means that it becomes possible to combine all the traditional function operations with the multiset operations.

For instance, it becomes possible and valid to write:

$$r(x) = (g(x) \cap h(x)) + g(x)$$
$$s(x) = (g(x) + h(x)) \cup (g(x) - h(x))$$
$$t(x) = [g(x) \cap h(x)] - [g(x) \cup h(x)]$$

where $r(x)$, $g(x)$ and $h(x)$ are generic real-valued functions on a support S.

These three functions are illustrated in Figure 3 assuming the function in Equation 74.

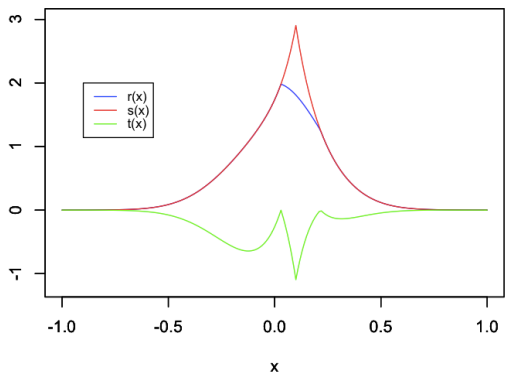


Figure 3: The functions $r(x)$, $s(x)$ and $t(x)$ obtained through mset operations.

We have already seen that the complement of a vector multset becomes the operation of sign change, and so we have with functions in the sense that $r(x)^C$ becomes $-r(x)$. This allows us to derive the following De Morgan extension to real-valued functions on a generic support S:

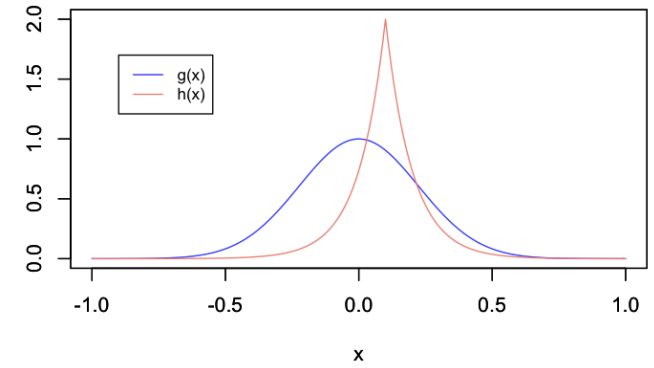
$$-[g(x) \cap h(x)] = -g(x) \cup -h(x)$$
$$-[g(x) \cup h(x)] = -g(x) \cap -h(x)$$

A similar approach can be applied to transform discrete scalar fields defined on more than one variables into respective msets, involving the index mapping described above.

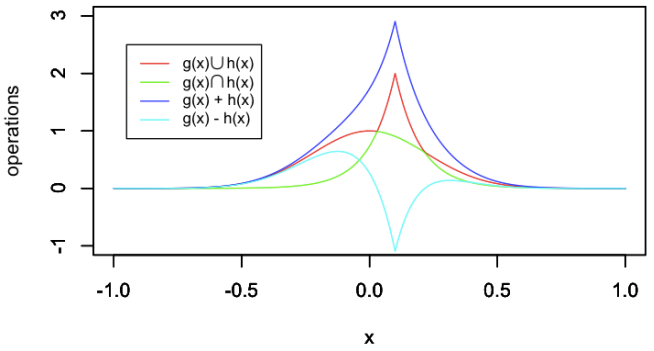
Let's illustrate the above concepts and possibilities in terms of the two following functions:

$$g(x) = e^{-10x^2}$$
$$h(x) = 2e^{-10|x-0.1|} \tag{74}$$

Figure 4 depicts the two above functions as well as their union, intersection, sum, and subtraction.



(a)



(b)

Figure 4: Two continuous functions $g(x)$ and $h(x)$ of a single variable (a), and their respective mset operations (b) of union, intersection, sum, and subtraction.

This examples illustrate several interesting points. First, we have the understanding of functions as respective msets immediately allows them to be operated by mset operations such as union and intersection. In addition, we observe that the sum of two functions is larger or equal than their union, as well as the possibility of the subtraction operation yielding negative values.

It is also interesting to observe the potential of the operations of union and intersection in producing sharp derivatives and discontinuities, which contributes an interesting manner of representing an ample range of function types as combinations of these operations, not to mention the operations of sum, subtraction, product and quotient.

Consequently, it becomes an interesting possibility to develop transformations of functions, analogous to the Fourier transform, considering not only series of basis functions, but also intersections and/or unions, and/or other possible hybrid operations between msets. One particularly useful benefit would be to become able to express functions with discontinuous derivatives as combinations of functions that are completely smooth. Also, it should be observed that the operations of sum and subtraction are bilinear, while the minimum and maximum, which mediate most of the mset operations, are not.

Another interesting perspective concerns in understanding the Jaccard similarity index from the perspective of real-valued msets, which yields:

$$\mathcal{J}_R(f(x), g(x)) = \frac{f(x) \sqcap g(x)}{f(x) \circledast g(x)} \tag{75}$$

where:

$$f(x) \circledast g(x) = \int_S \max(s_f f(x), s_g, g(x))dx \tag{76}$$

Related similarity indices have appeared previously in discrete manner in the context of analogies between the cosine similarity index in L1 spaces [30, 31].

Given that the Jaccard index does not take into account the relative interiority of the two compared sets [8], it has been combined with the interiority (or overlap e.g. [32]) index to yield the *coincidence index* between two functions, which can be expressed as:

$$\mathcal{C}_R(f(x), g(x)) = \mathcal{J}_R(f(x), g(x))\mathcal{I}_R(f(x), g(x)) \tag{77}$$

where:

$$\mathcal{I}_R(f(x), g(x)) = \frac{\int_S \min(s_f f(x), s_g, g(x))dx}{\min\{S_f, S_g\}} \tag{78}$$
$$S_f = \int_S |f(x)|dx \tag{79}$$
$$S_g = \int_S |g(x)|dx \tag{80}$$

The operation of *template matching* consists of, given a function $f(x)$, to quantify, along x, the similarity between its portions and another reference template function $g(x)$. This can be immediately implemented by by applying the real-valued Jaccard and coincidence indices to those two functions while one is slid respectively to the other, therefore implementing respective cross-correlations. High resulting values indicate portions of $f(x)$ that are similar to $g(x)$.

Figure 5 presents the result of matching the template in (b) with the function in (a) by using the traditional cross-correlation (c), the real-valued Jaccard index (d), the interiority index (e) and the and coincidence index (f).

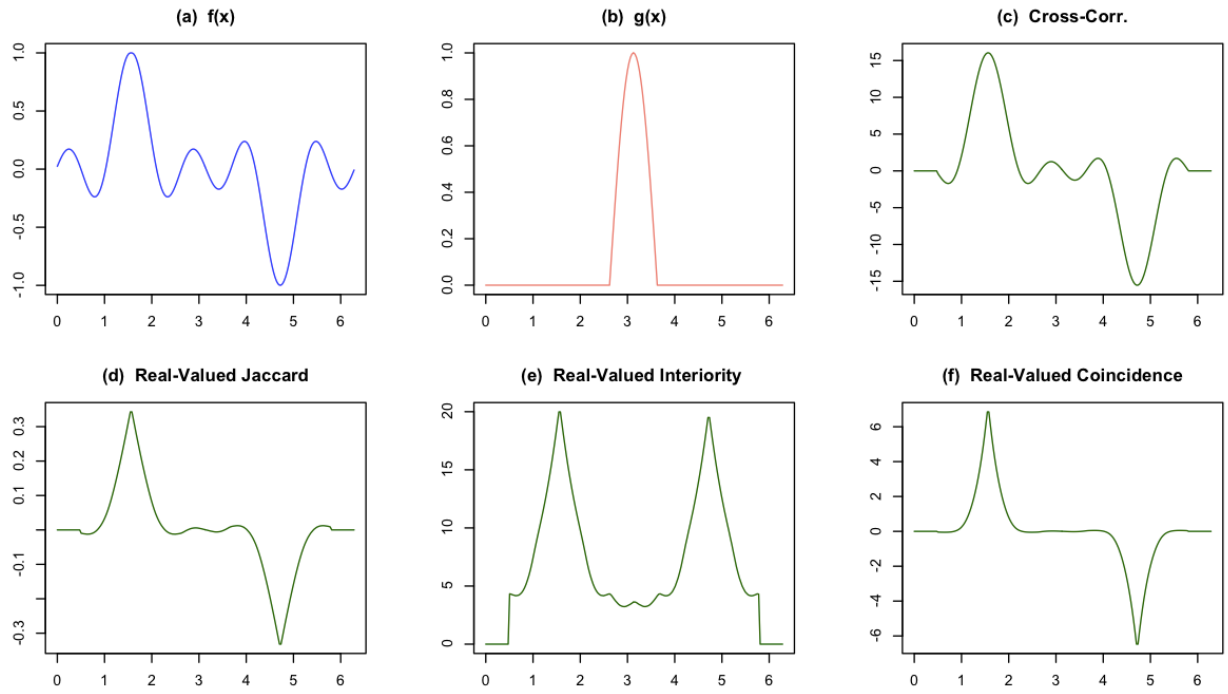


Figure 5: The templated in (b) is to be compared with the mfunction in (a) and the results obtained by using the traditional cross-correlation (c), the real-valued Jaccard index (d), the interiority index (e) and the and coincidence index (f). The potential of the real-valued Jaccard, and in particular the coincidence index, for obtaining narrower and sharper identification of the peak correspondence between the template and object function, while attenuating the effect of secondary matches, can be plainly observed.

As it can be verified, both the real-valued Jaccard and coincidence indices yielded a precise and well-localized identification of the maximum similarity between the portions of function $f(x)$ with the template function $g(x)$, not only for the positive parts, but also with respect to the negative. The secondary matchings appeared with substantially smaller values. This example illustrates the enhanced potential of the coincidence index for pattern recognition, filtering and neuronal network applications.

9 Multisets in Pattern Recognition

The possibility to use msets to represent any type of density paves the way to interesting applications in pattern recognition and deep learning (e.g. [9, 10, 33, 11, 12, 13]). In this section we illustrate how msets and the Jaccard index can be readily applied in order to quantify the similarity between two (or more) clusters represented by respective density functions, a frequent problem (e.g. [34]).

Let's consider the three sets of points in the scatterplot shown in Figure 6, which corresponds to the three species of iris flower in the frequently adopted iris dataset. Only two out of their 4 features have been chosen in the following example for simplicity's sake.

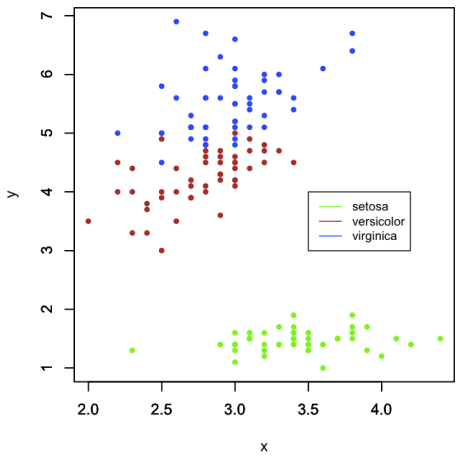


Figure 6: A scatterplot representing the distribution of three types of iris flowers represented by two respective features x and y.

The density obtained from the respective discrete samples through gaussian kernel expansion are shown in Figure 7.

The obtained multiset Jaccard index for each pairwise combination of categories are presented in Table 1.

The obtained results are fully compatible with the interrelationships between the three densities, or clusters, in Figure 6. In addition, the three-wise Jaccard index

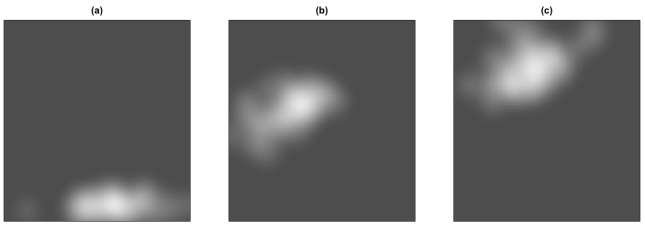


Figure 7: The three density scalar fields obtained by gaussian kernel expansion of each of the three categories.

	setosa	versicolor	virginica
setosa	1	2.6e-5	0
versicolor	2.6e-5	1	0.145
viginica	0	0.145	1

Table 1: The Jaccard indices obtained for pairwise combinations between the three iris species. The resolution has been limited to 6 digits.

from Equation 49 result nearly null, indicating a really small chance that the three densities correspond to the same cluster.

10 Concluding Remarks

The fascinating subject of msets has been presented in a hopefully introductory manner, followed by developments aimed at extending them to real-valued multiplicities. By allowing functions and fields to be understood as msets, several possibilities are made viable, some of which are explored in the present work.

In addition to introducing several of the basic mset concepts, the present work also proposed how the complement operation can be defined in a robust manner by allowing the subtraction of msets to take negative values. This paved the way for recovering several properties analogous to traditional sets involving the complement operation, including the De Morgan theorem, as well as to identifying additional mset operations.

The extension of msets to real functions and fields was also proposed, paving the way to defining functionals, of which the real-valued Jaccard and coincidence indices for are examples. The interpretation of several of the characteristics and properties of real-valued msets become particularly intuitive when approached in terms of functions. For instance, the fact that one function is contained into another can be graphically verified by checking the respective graphs. In addition, the incorporation of mset operations into real functions paves the way to obtained complex non-analytical functions (i.e. with derivative discontinuities) by combining smooth basic functions.

We have also illustrated the impressive potential of the coincidence index for performing template matching, yielding sharper and narrower detection peaks while attenuating secondary matches. These features are a consequence of the enhanced potential of the coincidence index in quantifying similarity in a detailed manner, which can be particularly useful in pattern recognition, deep learning, artificial intelligence and scientific modeling in general. For instance, neuronal networks can be constructed in which the synaptic input and integration, traditionally modeled and implemented in terms of the classic inner product, are performed by using the real-valued Jaccard or coincidence indices, therefore incorporating the respective advantages of these approaches.

The application of the real-valued Jaccard index to quantifying the similarity between two scalar fields corresponding to cluster densities of the iris dataset is also illustrated, with encouraging results.

The presented concepts and methods pave the way to several interesting applications, also motivating further integrations between the structures and properties between the domains of set theory, propositional logic, and analysis.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

[1] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[2] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[3] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[4] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[5] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[6] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.

[7] L. da F. Costa. An ample approach to modeling. Researchgate, 2019. https://www.researchgate.net/

publication/355056285_An_Ample_Approach_to_Data_and_Modeling. [Online; accessed 10-Oct-2021.].

[8] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[10] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[11] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[12] J. Schmidhuber. Deep learning in neural networks:an overview. *Neural networks*, 61:85–117, 2015.

[13] H. F. de Arruda, A. Benatti, C. H. Comin, and L. da F. Costa. Learning deep learning. Researchgate, 2019. https://www.researchgate.net/publication/335798012_Learning_Deep_Learning_CDT-15. [Online; accessed 22-Dec-2019.].

[14] L. da F. Costa. Modeling: The human approach to science. Researchgate, 2019. https://www.researchgate.net/publication/333389500_Modeling_The_Human_Approach_to_Science_CDT-8. [Online; accessed 1-Oct-2020.].

[15] W. Rudin. *Elemets of Algebraic Topology*. Addison-Wesley, 1984.

[16] W. D. Blizard. Real-valued multisets and fuzzy sets. *Fuzzy Sets and Systems*, 33:77–97, 1989.

[17] W. D. Blizard. Negative membership. *Notre Dame Journal of Formal Logic*, 31:346–368, 1990.

[18] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société vaudoise des sciences naturelles*, 37:241–272, 1901.

[19] A. Schubert. Measuring the similarity between the reference and citation distributions of journals. *Scientometrics*, 96:305–313, 2013.

[20] A. Schubert and A. Telcs. A note on the jaccardized czekanowski similarity index. *Scientometrics*, 98:1397–1399, 2014.

[21] O Rozinek and J. Mareš. The duality of similarity and metric spaces. *Appl. Sciences*, 11:10.3390, 2021.

[22] M Ružička. Anwendung mathematisch-statistischer methoden in der geobotanik. *Biológica*, 13:647–661, 1958.

[23] Wikipedia. Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index. [Online; accessed 10-Oct-2021].

[24] B. K. Samanthula and W. Jiang. Secure multiset intersection cardinality and its application to jaccard coefficient. *IEEE Transactions on Dependable and Secure Computing*, 13(5):591–604, 1989.

[25] D. Bacciu, A. Micheli, and A. Sperduti. Generative kernels for tree-structured data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4932–4946, 2018.

[26] Y. Yuan, M. Chao, and Y.-C. Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, 36(9):1876–1886, 2017.

[27] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Ketters, H. Rousseau, and A. Vanhoutte. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–318, 1989.

[28] L. Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

[29] S. Park and D.-Y. Kim. Assessing language discrepancies between travelers and online travel recommendation systems: Application of the jaccard distance score to web data mining. *Technological Forecasting and Social Change*, 123:381–388, 2017.

[30] C. E. Akbas, A. Bozkurt, M. T. Arslan, H. Aslanoglu, and A. E. Cetin. L1 norm based multiplication-free cosine similiarity measures for big data analysis. In *IEEE Computational Intelligence for Multimedia Understanding (IWCIM)*, France, Nov. 2014.

[31] C. E. Akbas, A. Bozkurt, A. E. Cetin, R. Cetin-Atalay, and A. Uner. Multiplication-free neural networks. In *Signal Processing and Communications Applications Conference (SIU)*, Malatya, Turkey, May. 2015.

[32] M. K. Vijaymeena and K. Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications*, 3(1):19–28, 2016.

[33] R. J. Schalkoff. *Digital Image Processing and Computer Vision*. Wiley, 1989.

[34] J. D. Loudin and H. E Miettinen. A multivariate method for comparing n-dimensional distributions. In *PHYSTAT2003, SLAC*, 2003.