https://doi.org/10.1016/j.fss.2019.04.018

Granular Counting of Uncertain Data

C. Mencar[a,*], W. Pedrycz[b,c],

[a] *Department of Informatics, University of Bari "A. Moro", Bari, Italy*
[b] *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada.*
[c] *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Abstract

We propose a definition of granular count realized in the presence of uncertain data modeled through possibility distributions. We show that the resulting counts are fuzzy intervals in the domain of natural numbers. Based on this result, we devise two algorithms for granular counting: an exact counting algorithm with quadratic-time complexity and an approximate counting algorithm with linear-time complexity. We compare the two algorithms on synthetic data and show their application to a Bioinformatics scenario concerning the assessment of gene expressions in cells.

Keywords: Granular Computing, Possibility Theory, Counting, Fuzzy Numbers

1. Introduction

Data collection is a common practice that is preliminary to any analysis. In essence, data are collected by a systematic application of measurement of a target variable, in order to discover unknown relations, testing hypotheses, etc. Modern technology allows to collect huge amounts of data, which call for complex methodologies for understanding and analyzing data. Data Science

*Corresponding author
Email addresses: corrado.mencar@uniba.it (C. Mencar), wpedrycz@ualberta.ca (W. Pedrycz)

is an interdisciplinary field devoted to the study of these technologies, and its interest in the scientific community is incessantly rising.

Depending on the application, data can be either raw or derived [1]. In particular, derived data are produced through additional processing or analysis of raw data. Both raw and derived data may be contaminated by uncertainty, i.e. a lack of confidence about the exact value of the observed variable. There are several reasons for uncertainty present in data [2]; reasonably, the more complex is the phenomenon under investigation, the more likely collected data are uncertain.

There are at least four strategies to deal with uncertain data: *understand*, *minimize*, *exploit* and *ignore* uncertainty [3]. The simplest one is just to ignore uncertainty, but this strategy is dangerous as it may introduce bias in the subsequent processing stages, which is hard to recognize. Minimizing uncertainty is important but, in most cases, a residual uncertainty is still present in data and further analysis must cope with it. Uncertainty can be exploited by propagating it in the subsequent data processing stages. In this way, the results of data analysis show their uncertainty which can be assessed in order to judge their final utility. But, in order to be exploited, uncertainty must be understood and modeled by a proper theoretical framework.

Commonly, a probabilistic framework is used to model uncertainty [2]. There is an extensive literature of data analysis and mining methods dealing with uncertain data modeled in a probabilistic framework, such as machine learning with noisy labels [4, 5, 6, 7], data indexing [8], uncertain data mining [9], query ranking [10], etc. However, at the same time there are further studies that cope with uncertainty in data with different frameworks, such as classical sets [11], rough sets [12], evidence theory [13] and possibility theory [14].

Possibility theory is the simplest mathematical theory dealing with uncer-

tainty due to incomplete information [15]. It has been used to represent and process data uncertainty in several scenarios, including information fusion [16], statistical reasoning [15], supervised machine learning, where uncertainty can be either in the observed data [17] or in the target variables [18], or both [19]. Possibility theory offers an interpretation of fuzzy sets [20], thus it serves as a theoretical underpinning for fuzzy data analysis [21].

A basic operation with data is counting, i.e. finding the number of data samples having a specific value. Counting is often a preliminary step for several types of analysis, such as descriptive statistics, comparisons, etc. This is quite a simple operation if objects are represented as precise data[1], but it becomes non-trivial when data are uncertain. In fact, uncertainty in data must propagate in counting, therefore results are *granular* rather than precise.

A first approach to counting with uncertain data is to remove uncertainty prior to counting. A typical way to remove uncertainty in data is to select just one value that is judged most appropriate according to some criteria (e.g., the most probable). This is a very common approach indeed, but it introduces bias due to the arbitrary removal of alternative values data could assume. As a result, precise counts can be computed (with their subsequent uses) but the value of the results can be hampered by an arbitrary choice. However, propagating uncertainty in counting could be very complex, especially if uncertainty in data is represented in terms of probability distributions of values.

Counting data resembles the computation of the cardinality of a set. In fact, the cardinality of a set is informally defined as the number of elements belonging to the set. If a set includes all data samples with a specific value, counting all data with that value coincides with the cardinality of the corresponding set.

[1]The complexity of counting also depends on the procedure that is required to recognize that an object belongs to a set. However, here we do not take into account such complexity.

Therefore, since uncertain data modeled with Possibility Theory give rise to fuzzy sets of data samples, a first choice would be to use one of the available methods to compute the cardinality of a fuzzy set [22, 23, 24, 25].However, there is a profound difference between counting (possibilistic) uncertain data and the cardinality of a fuzzy set. This difference manifests in the different semantics that are assumed. On the one hand, the cardinality of a fuzzy set makes sense when it is assumed that *all* elements actually belong to the set (to some degree). On the other hand, uncertain data have just one value (thus they belong to one set only) although it is not generally known; what is known is a possibility distribution of values. Therefore, there is no certainty that a data sample belongs to a set (corresponding to a value), hence cardinality methods do not provide any information about the number of data samples with a specific value, i.e. they do not count data.

Counting uncertain data calls for a completely different approach that is described in this paper. After some preliminary definitions (Sec. 2), we introduce the concept of granular counting (Sec. 3) and show that a granular count of possibilistic data is a fuzzy interval. Based on this, in Sec. 4 we develop two algorithms for granular counting: one for exact counting and another for approximate but fast counting. In the experimental section (Sec. 5) we first report a number of experiments on synthetic data that quantify the differences between exact and approximate counting; then we illustrate a case study in Bioinformatics where the usefulness of granular counting is shown in a real-world scenario. The paper ends with some conclusive notes and directions for future research.

2. Preliminaries

Since we want to deal with primary and derived data seamlessly, we introduce a specific terminology. We assume the existence of a collection of objects

or REFERENTS, which are detected through OBSERVATIONS. It is important to notice that the effect of an observation is a data sample that can be used as is, or further processed, in order to refer to one of the referents. (In short, we say that an observation refers to a referent.) It is also important to underline the relation between observation and referent — which is called *reference* — because we introduce uncertainty in this relation. There might be several reasons for such uncertainty; on a very general level, we assume that uncertainty is due to incomplete information coming from an observation, which impedes a unequivocal reference to one of the referents. We model such uncertainty with Possibility Theory [14].

Given a set R of referents, a possibility distribution is a mapping from R to a totally ordered scale. We use the interval $[0, 1]$ for such a scale, therefore a possibility distribution is a function

$$\pi : R \mapsto [0, 1]$$

such that $\exists r \in R : \pi(r) = 1$.

A possibility distribution represents a state of knowledge about the reference relation of an observation to a referent. For an observation with possibility distribution π the value $\pi(r) = 0$ means that it is impossible that the referent r is referred by the observation, while $\pi(r) = 1$ means that the referent r is absolutely possible (though not certain). Intermediate values of $\pi(r)$ stand for gradual values of possibility, which quantifies the completeness of information resulting from an observation.

It is important to underline that a possibility distribution quantifies completeness of negative information, i.e. the lower the possibility value, the more information we have to *exclude* a referent. Even in the case of maximal possibility, we may still have not the certainty that observation refers to a particular

referent. For example, if $\pi(r_1) = \pi(r_2) = 1$ with $r_1 \neq r_2$, it is fully possible that the observation refers to either r_1 or r_2 (in other words, the observation is ambiguous), but if $\pi(r_1) = 1$ and $\pi(r) = 0$ for $r \neq r_1$ we conclude that it is possible that the observation refers to r_1 but it is impossible that the observation refers to any other referent; in other words we infer that the observation certainly refers to r_1.

In the case of multiple observations, we can combine the possibility values through appropriate operators. In particular, we assume that the observations are *unrelated* (or non-interactive), meaning that the real referent of an observation does not influence the real referent of another observation. In this case, we are allowed to combine the possibility distributions of n observations which are aggregated through the minimum function [20]:

$$\pi\left(\left[r^{(1)}, r^{(2)}, \ldots, r^{(n)}\right]\right) = \min_i \pi_{o_i}\left(r^{(i)}\right) \tag{1}$$

being π_{o_i} the possibility distribution assigned to observation o_i. In the binary case, the left-hand part of (1) is maximal if, for each observation o_i, it is possible that o_i refers to $r^{(i)}$ and is null if for at least one observation o_i, it is impossible that it refers to $r^{(i)}$.

On the other hand, the possibility degree that an observation refers to a referent in a subset $S \subseteq R$ is determined by the maximum possibility degree of all the referents in the subset:

$$\pi(S) = \max_{r \in S} \pi(r) \tag{2}$$

In the binary case, $\pi(S)$ is maximal if there exists at least one reference $r \in S$ that is possibly referred by the observation, and is minimal if all referents in S are impossible.

3. Granular counts

Based on the preliminary definition, we can define a granular count of observations for a given referent. After that, we provide two algorithms for an effective computation of granular counts.

3.1. Definition of granular count

Let R be a finite, non-empty set of objects or *referents*:

$$R = \{r_1, r_2, \ldots, r_n\} \tag{3}$$

and let O be a finite, non-empty set of *observations*[2]:

$$O = \{o_1, o_2, \ldots, o_m\} \tag{4}$$

Let $\pi_{o_j}(r_i) \in [0, 1]$ be the possibility degree that observation o_j refers to referent r_i. Therefore, an observation actually refers to one referent only, but this association is uncertain. The following question is put forward: *how many observations refer to each referent*[3]*?*

For the sake of simplicity, the binary case is first assumed, i.e. $\pi_{o_j}(r_i) \in \{0, 1\}$ for each observation and referent. Let N_i be the number of observations associated to referent r_i. Some special cases can be easily recognized:

- $\forall j : \pi_{o_j}(r_i) = 0$, i.e., it is impossible that observation o_j refers to r_i, for every o_j. In such a case, certainly $N_i = 0$;
- Even if, for some j, $\pi_{o_j}(r_i) = 1$, it is still possible that $N_i = 0$ if, for each observation o_j, it is possible that o_j refers to another referent $r_{i'} \neq r_i$;

[2]We will assume the observations as non-interacting, i.e. the value assumed by an observation does not influence the value assumed by another observation.

[3]We assume that the set R consists of the only referents that each observation can refer to.

- If there exists j such that $\pi_{o_j}(r_i) = 1$ and, for all other indexes $i' \neq i$, it is $\pi_{o_j}(r_{i'}) = 0$ then the observation o_j *must* refer to r_i. It is therefore impossible that $N_i = 0$.

These considerations highlight the necessity of defining a possibility distribution on N_i, which will be denoted by π_{N_i}. (Therefore, the count of observation is granular rather than precise.) The value $\pi_{N_i}(x)$ is the possibility degree that *exactly* x observations refer to r_i.

Informally speaking, it is possible that exactly x observations refer to r_i if there exists a subset of x observations such that it is possible that all and only these observations refer to r_i. In order to formalize this statement we consider the class of functions

$$a : O \mapsto R$$

that assign observations $o \in O$ to referents $r \in R$. An assignment a is *admissible* with respect to the possibility distributions π_o iff, for all $o \in O$, $a(o) = r$ implies $\pi_o(r) > 0$.

Let $O_x \subseteq O$ be a subset of x observations, i.e. $|O_x| = x$. (The number of such subsets is $\binom{m}{x}$ for $x \leq m$, and 0 otherwise; for $x = 0$ the unique possible subset is $O_0 = \emptyset$.) The possibility that *all and only* the elements of O_x refer to r_i is given by the admissibility of any function taking values as follows:

$$a(o) = \begin{cases} r_i & \text{if } o \in O_x \\ r \neq r_i & \text{if } o \notin O_x \end{cases}$$

As a consequence, the possibility that all and only the elements of O_x refer to r_i equates the possibility that all elements of O_x refer to r_i and all the remaining elements in $O \setminus O_x$ refer to other referents. In the binary case, we can formalize

this property as follows:

$$(\forall o \in O_x : \pi_o(r_i) = 1) \wedge (\forall o \notin O_x : \exists r \neq r_i : \pi_o(r) = 1) \tag{5}$$

which ensures that all observations in the subset O_x possibly refer to r_i and that any observation outside O_x possibly refers to at least another referent $r \neq r_i$. It is important to observe that this does not prevent an observation $o' \notin O$ to possibly refer to r_i if $\pi_{o'}(r_i) = 1$ (as well as to another reference $r \neq r_i$, i.e. $\pi_{o'}(r) = 1$). Even in this case, however, it is plainly possible that O_x is the set of all and only observations referring to r_i since it is possible that o' actually refers to $r \neq r_i$ and not to r_i. The truth of proposition (5) does not ensure that O_x is the *actual* subset of all observations referring to r_i, but only that O_x is a *possible* subset of all such observations.

It could be interesting to compare eq. (5) with a perhaps more intuitive definition, which, however, does not capture the intended meaning of granular count:

$$(\forall o \in O_x : \pi_o(r_i) = 1) \wedge (\forall o \notin O_x : \pi_o(r_i) = 0) \tag{5'}$$

Eq. (5') states that the possibility that all and only the observations in O_x refer to r_i is satisfied when it is possible that all observations on O_x refer to r_i and for all observations outside O_x it is impossible that they refer to r_i. However, this is not a correct definition because it excludes the case that an observation o' outside O_x possibly refers to r_i as well as to another referent $r \neq r_i$. In fact, in such a case, proposition (5') is false even if it is plainly possible that O_x is the set of all and only observations of r_i provided that o' actually refers to r.Since we are interested in defining a counting method that takes into account graded possibility, we express property (5) by using the possibility operators

and the assumption of non-interactivity, which yield the following definition of possibility:

$$\pi_{O_x}(r_i) = \min\left\{\min_{o\in O_x}\pi_o(r_i), \min_{o\notin O_x}\max_{r\neq r_i}\pi_o(r)\right\} \tag{6}$$

with the convention that $\min\emptyset = 1$. Eq. (6) defines the possibility degree that O_x is the subset of all and only the observations of r_i by computing the least possibility degree of two simultaneous events: (i) all observations of O_x refer to r_i, and (ii) all the other observations refer to a different referent. To compute the possibility of event (i), since observations are assumed to be unrelated, the minimum possibility degree among all observations in O_x is computed; on the other hand, to compute the possibility degree of event (ii), for all observations outside O_x, the possibility degree that each of them refers to any referent $r \neq r_i$ is computed (this is easily accomplished by computing the maximum possibility degree of each observation over all referents different from r_i), then the minimum degree is retained.

Example 1. A simple example may clarify definition (6). Given the following table:

	r_1	r_2
o_1	1	0
o_2	1	1
o_3	1	1
o_4	0	1

we could ask whether it is possible that no observation refers to r_1. We immediately observe that this is impossible: the observation o_1 can refer to r_1 only. In fact, if we consider the (unique) set $O_0 = \emptyset$, by applying (6) we obtain

$$\pi_{O_0}(r_1) = \min\left\{\min_{o\in O_0}\pi_o(r_i), \min_{o\notin O_0}\pi_o(r_2)\right\} = \min\left\{1, \min_{o\in O}\pi_o(r_2)\right\} = 0$$

Now we consider $O_1 = \{o_1\}$. Apparently, it is possible that o_1 in the unique observation of r_1 since o_2 e o_3 could refer to r_2 while o_4 cannot refer to r_1. In fact

$$\pi_{O_1}(r_1) = \min\{\pi_{o_1}(r_1), \min\{\pi_{o_2}(r_2), \pi_{o_3}(r_2), \pi_{o_4}(r_2)\}\} = 1$$

On the other hand, if $O_1 = \{o_2\}$, then

$$\pi_{O_1}(r_1) = \min\{\pi_{o_2}(r_1), \min\{\pi_{o_1}(r_2), \pi_{o_3}(r_2), \pi_{o_4}(r_2)\}\} = 0$$

because $\pi_{o_1}(r_2) = 0$. This means that, as it is evident from the table, o_2 cannot be the unique observation of r_1. We will achieve the same result for $O_1 = \{o_3\}$, while, trivially, $\pi_{O_1}(r_1) = 0$ when $O_1 = \{o_4\}$.

We can now proceed to consider sets of two observations (there are $\binom{4}{2} = 6$ different sets). We first consider $O_2 = \{o_1, o_2\}$. It is evidently possible that only these two observations can be associated to r_1. Indeed:

$$\pi_{O_2}(r_1) = \min\{\min\{\pi_{o_1}(r_1), \pi_{o_2}(r_1)\}, \min\{\pi_{o_3}(r_2), \pi_{o_4}(r_2)\}\} = 1$$

We notice that $o_3 \notin O_2$ yet it is possible that o_3 refers to r_1 since $\pi_{o_3}(r_1) = 1$. However, we also notice that $\pi_{o_3}(r_2) = 1$, therefore, the situation where o_1 and o_2 are the only actual observations referring to r_1 while both o_3 and o_4 refer to r_2 cannot be excluded. In other words, it is possible (though not necessarily true) that O_2 is the subset of all and only observations referring to r_1. Finally, it should be noticed that proposition (5') is false because it is possible—*although not necessary*—that observation o_3 refers to referent r_1, but this does not prevent the *possibility* that $\{o_1, o_2\}$ are the only observations that refer to r_1. Actually, expression (5') requires that observations in O_x can refer to referent r_i and observations outside O_x cannot refer to r_i. This is, however, a stronger requirement than (5) which does not reflect the concept

of admissible assignment: in fact, turning back to the example, according to (5') it is impossible that r_1 can be referred by two observations only, but this contradicts the semantics of the possibility matrix π; according to the matrix π, there are two sets of two observations that can possibly be the sole observations of r_1, namely $O_2 = \{o_1, o_2\}$ and $O_2' = \{o_1, o_3\}$. (Notice that $O_2'' = \{o_2, o_3\}$ cannot be such a set because o_1 *must* refer to r_1.) If we apply (5) to these sets we obtain, respectively, true for O_2 and O_2' and false for O_2''.

The same considerations apply to $O_2 = \{o_1, o_3\}$ but not for $O_2 = \{o_1, o_4\}$ because $\pi_{o_4}(r_1) = 0$. It is noteworthy to observe that for $O_2 = \{o_2, o_3\}$ we obtain

$$\pi_{O_2}(r_1) = \min\{\min\{\pi_{o_2}(r_1), \pi_{o_3}(r_1)\}, \min\{\pi_{o_1}(r_2), \pi_{o_4}(r_2)\}\} = 0$$

because $\pi_{o_1}(r_2) = 0$. In fact, although it is possible that o_2 e o_3 can refer to r_1, these cannot be the sole observations associated to r_1, since certainly o_1 must be included too. For similar reasons, we have $\pi_{O_2}(r_1) = 0$ for $O_2 = \{o_2, o_4\}$ and $O_2 = \{o_3, o_4\}$.

We can go further on with sets of three observations. For $O_3 = \{o_1, o_2, o_3\}$ we obtain, as expected:

$$\pi_{O_3}(r_1) = \min\{\min\{\pi_{o_1}(r_1), \pi_{o_2}(r_1), \pi_{o_3}(r_1)\}, \min\{\pi_{o_4}(r_2)\}\} = 1$$

while in all the other cases, i.e. $O_3 = \{o_1, o_3, o_4\}$, $O_3 = \{o_2, o_3, o_4\}$, $O_3 = \{o_1, o_2, o_4\}$ we have $\pi_{O_3}(r_1) = 0$ because of the presence of o_4 in O_3. In any case, there exists the possibility that three observations refer to r_1, while it is impossible that we have four. In fact, for $O_4 = O$ we have

$$\pi_{O_4}(r_1) = \min\{\pi_{o_1}(r_1), \pi_{o_2}(r_1), \pi_{o_3}(r_1), \pi_{o_4}(r_1)\} = 0$$

This example shows that it is possible to compute the possibility that a specific set of x observations collects the sole observations associated to a referent. However, we are not interested in a specific set, but in *any set* of x elements. We can therefore define the possibility value that the number of observations for a referent r_i is x as:

$$\pi_{N_i}(x) = \max_{O_x \subseteq O} \pi_{O_x}(r_i) \tag{7}$$

for $x \leq m$ and $\pi_{N_i}(x) = 0$ for $x > m$. Eq. (7) provides a granular definition of count. Counting is imprecise because observations are uncertain.

By reconsidering the previous example, we have:

$$\pi_{N_1}(x) = \begin{cases} 0, & x = 0 \\ 1, & 1 \leq x \leq 3 \\ 0, & x = 4 \end{cases}$$

i.e., the uncertainty of associating a referent to each observation translates into a set of possible counting values. For binary possibility, the result of counting is a crisp set. Accordingly, gradual uncertainty yields a fuzzy set of counts.

Example 2. In order to show how graded possibility leads to a fuzzy set of counts, we may consider the following table:

	r_1	r_2	r_3
o_1	1	0	0
o_2	1	.8	.6
o_3	1	1	0
o_4	.6	.8	1
o_5	.5	1	0

and focus on r_1. We immediately observe that $\pi_{N_1}(0) = 0$ since o_1 must refer to r_1: at least one observation must be associated to the referent. If we consider $O_1 = \{o_1\}$, by applying (6) we obtain

$$\begin{aligned}
\pi_{O_1}(r_1) &= \min\{\pi_{o_1}(r_1), \min\{\max\{\pi_{o_2}(r_2), \pi_{o_2}(r_3)\}, \ldots \max\{\pi_{o_5}(r_2), \pi_{o_5}(r_3)\}\}\} \\
&= \min\{\max\{.8, .6\}, \ldots \max\{1, 0\}\} \\
&= \min\{.8, 1, 1, 1\} \\
&= .8
\end{aligned}$$

it is therefore possible that o_1 is the sole observation of r_1 but this possibility is attenuated by the fact that, for at least another observation (o_2), the possibility of being associated to another referent is not maximal (therefore it is more possible that it is associated to r_1). As a consequence, it is more possible that both o_1 and o_2 refer to r_1 than the case that o_1 only refers to r_1: this is quantified by a reduced possibility degree that o_1 is the only observation of r_1. For any other instance of $O_1 \neq \{o_1\}$ we always have $\pi_{O_1}(r_1) = 0$ because o_1 can only be associated to r_1. From (7) it follows that $\pi_{N_1}(1) = 0.8$.

We can now consider sets of two observations. Obviously, for all sets that do not include o_1 we have $\pi_{O_2}(r_1) = 0$. We therefore take the case $O_2 = \{o_1, o_2\}$. Then

$$\begin{aligned}
\pi_{O_2}(r_1) &= \min\{\min\{\pi_{o_1}(r_1), \pi_{o_2}(r_1)\}, \min\{\max\{\pi_{o_3}(r_2), \pi_{o_3}(r_3)\}, \ldots \max\{\pi_{o_5}(r_2), \pi_{o_5}(r_3)\}\}\} \\
&= \min\{1, \min\{\max\{1, 0\}, \max\{.8, 1\}, \max\{1, 0\}\}\} \\
&= \min\{1, 1, 1\} \\
&= 1
\end{aligned}$$

since, for all observations *not* in O_2, there exists maximum possibility that they

refer to other referents different from r_1. It is worthless to carry on calculations, because it is immediately deducible that $\pi_{N_1}(2) = 1$.

Likewise toO_2, for any O_3 we can easily deduce $\pi_{O_3}(r_1) = 0$ when $o_1 \notin O_3$. Furthermore, for $O_3 = \{o_1, o_2, o_3\}$ we obtain $\pi_{O_3}(r_1) = 1$, hence $\pi_{N_1}(3) = 1$. For O_4 we must observe that there are not four observations with maximum possibility to r_1. In fact, the maximum possibility degree is achieved with $O_4 = \{o_1, o_2, o_3, o_4\}$ equal to

$$\begin{aligned} \pi_{O_4}(r_1) &= \min\{\min\{\pi_{o_1}(r_1), \ldots, \pi_{o_4}(r_1)\}, \max\{\pi_{o_5}(r_2), \pi_{o_5}(r_3)\}\} \\ &= \min\{.6, 1\} \\ &= .6 \end{aligned}$$

Therefore, $\pi_{N_1}(4) = 0.6$, i.e. the possibility that four observations refer to the same referent r_1 is not high. Even smaller, though not null, is the possibility $\pi_{N_1}(5) = 0.5$. In fact, observations o_4 and o_5 are more likely to refer to referents different from r_1.

3.1.1. Relations with Formal Concept Analysis

It is interesting to observe a parallel with formal concept analysis (FCA), which develops around the notion of formal context, i.e. a tuple (G, M, I) where G is a collection of objects, M is a collection of properties (or attributes) and $I \subseteq G \times M$ represents a relation that can be read as $gIm \equiv$ 'object g has property m' [26]. Based on a formal context, a possibility distribution $\pi_m(g)$ could be defined on the set G of objects so as to express the possibility that an unknown object is $g \in G$ when a property $m \in M$ is observed [27]. Based on this possibility distribution, a number of set functions can be defined, which enable a formal characterization of all possible relations between objects and properties. In this setting, a parallel can be established by treating referents

as objects ($G = R$) and observations as properties ($M = O$); in this way it is possible to compare the possibility distributions $\pi_m(g)$ in formal contexts and the possibility distributions $\pi_{o_j}\left(r^{(i)}\right)$ as defined in this paper.

In FCA, a possibility distribution $\pi_m(g)$ is determined by the relation I, i.e. $\pi_m(g) = 1$ iff object g has property m, therefore, in order to count how many properties an object holds, the calculation is trivial: it is sufficient to sum all possibility degrees by varying m. This is possible because a property m is shared among different objects: if an object g has a property m, this does not prevent another object $g' \neq g$ to have the same property m. (We could formally express this situation by writing $\pi_m(g) = \pi_m(g|g')$, i.e., the possibility that an unknown object is g given an observed property m does not depend by the knowledge that another object g' has the same property.) On the other hand, in the context of this paper the count of observations referring to a referent $r^{(i)}$ leads to the more complex formulation reported in eq. (7), because an observation refers to one referent only and is not, unlike properties, shared among other referents. (We could formally express this situation by writing $\pi_{o_j}\left(r^{(i)}|r^{(k)}\right) = 0$ for $i \neq k$, i.e., it is impossible that observation o_i refers to referent $r^{(i)}$given that it refers to $r^{(k)}$ for $k \neq i$.)

The possibility distributions developed within FCA are defined from different grounds than those used in this paper;this is why we used a different nomenclature from FCA. On the other hand, $\pi_{o_i}\left(r^{(i)}\right)$ can be interpreted as the possibility that an object o_i takes value $r^{(i)}$ in an information system characterized by a single attribute a with domain $U_a = R$ [28]. However, the restriction to a single attribute makes the formalism of information systems impractical for the developments of this paper; nevertheless, the proposed method shows a way of counting the number of objects possessing an attribute value in the context of information systems characterized by possibilistic uncertainty.

3.2. Properties of granular count

In this section, we prove that a granular count is a fuzzy interval in the domain of natural numbers. A fuzzy interval is a convex and normal fuzzy set on a numerical domain (in our case, it is $\mathbb{N}$). Convexity of a fuzzy set can be established by proving that all α-cuts are intervals, while normality is guaranteed if at least one element of the domain has full membership in the fuzzy set.

From (6) we observe that, when we focus on a specific referent r_i, all the other referents are aggregated through the max operator. Thus, we can simplify the notation by dropping the index i and introducing a virtual referent $\bar{r}$ such that

$$\pi_o(\bar{r}) = \max_{r' \neq r} \pi_o(r') \tag{8}$$

In this way, (6) can be simplified as:

$$\pi_{O_x}(r) = \min\left\{ \min_{o \in O_x} \pi_o(r), \min_{o \notin O_x} \pi_o(\bar{r}) \right\} \tag{9}$$

Also, (7) can be simplified as:

$$\pi_N(x) = \max_{O_x \subseteq O} \pi_{O_x}(r) \tag{10}$$

3.2.1. Normality of granular count

It is easy to observe that π_N is a fuzzy set defined on the domain of natural numbers:

$$\pi_N : \mathbb{N} \mapsto [0, 1]$$

we can, therefore, consider α-cuts of π_N, i.e.

$$[\pi_N]_\alpha = \{x \in \mathbb{N} | \pi_N(x) \geq \alpha\}$$

In particular, we can consider the core of π_N, namely

$$\operatorname{core}(\pi_N) = [\pi_N]_1 = \{x \in \mathbb{N} | \pi_N(x) = 1\}$$

A question concerning the core is if it could be empty. From the definition, we have:

$$x \in \operatorname{core}(\pi_N) \Leftrightarrow \pi_N(x) = 1 \Leftrightarrow \max_{O_x \subseteq O} \pi_{O_x}(r) = 1$$

The last condition is verified if and only if

$$\exists O_x \subseteq O : \pi_{O_x}(r) = 1$$

which is equivalent to

$$\exists O_x \subseteq O : \min\left\{\min_{o \in O_x} \pi_o(r), \min_{o \notin O_x} \pi_o(\bar{r})\right\} = 1$$

that is:

$$\min_{o \in O_x} \pi_o(r) = 1 \wedge \min_{o \notin O_x} \pi_o(\bar{r}) = 1$$

If we consider that each π_o is a (normal) possibility distribution, then, for each $o \in O$:

$$\max_{r \in R} \pi_o(r) = 1 \tag{11}$$

that is:

$$\max\left\{\pi_o(r), \max_{r' \in R \setminus \{r\}} \pi_o(r')\right\} = \max\{\pi_o(r), \pi_o(\bar{r})\} = 1$$

i.e., for each $o \in O$, $\pi_o(r) = 1$ or $\pi_o(\bar{r}) = 1$ (or both).

Let $O^{(1)} = \{o \in O | \pi_o(r) = 1\}$ ($O^{(1)}$ could possibly be empty) and $\bar{O}^{(1)} =$

$O \backslash O^{(1)}$. Therefore

$$\forall o \notin O^{(1)} : \pi_o(r) < 1$$

but, because of normality (11),

$$\forall o \notin O^{(1)} : \pi_o(\bar{r}) = 1$$

therefore

$$\min\left\{ \min_{o \in O^{(1)}} \pi_o(r), \min_{o \notin O^{(1)}} \pi_o(\bar{r}) \right\} = 1$$

thus, if $x^{(1)} = \left|O^{(1)}\right|$, then $\pi_N\left(x^{(1)}\right) = 1$, therefore the core of π_N cannot be empty. This allows us to state that π_N is normal.

3.2.2. Convexity of granular count

Because of normality, we can claim that any α-cut of π_N is not empty. (Indeed, for any fuzzy set A, $[A]_\alpha \supseteq [A]_\beta$ if $\alpha \leq \beta$; therefore $[\pi_N]_\alpha \supseteq [\pi_N]_1 = \text{core}(\pi_N) \neq \emptyset$.) We are ready to extend our argument to any α-cut. Let $0 < \alpha \leq 1$; then $x \in [\pi_N]_\alpha$ if and only if

$$\max_{O_x \subseteq O} \pi_{O_x}(r) \geq \alpha$$

i.e.

$$\exists O_x \subseteq O : \pi_{O_x}(r) \geq \alpha$$

Since $[\pi_N]_\alpha$ is non-empty, for each $x \in [\pi_N]_\alpha$ the existence of (at least one) O_x is verified. By definition:

$$\pi_{O_x}(r) = \min\left\{ \min_{o \in O_x} \pi_o(r), \min_{o \notin O_x} \pi_o(\bar{r}) \right\} \geq \alpha$$

that is

$$\min_{o\in O_x} \pi_o(r) \geq \alpha \wedge \min_{o\notin O_x} \pi_o(\bar{r}) \geq \alpha$$

i.e.

$$\forall o \in O_x : \pi_o(r) \geq \alpha \wedge \forall o \notin O_x : \pi_o(\bar{r}) \geq \alpha$$

which is true for each $x \in [\pi_N]_\alpha$.

We define:

$$O_{\min} = \{o \in O | \pi_o(r) \geq \alpha \wedge \pi_o(\bar{r}) < \alpha\} \tag{12}$$

and

$$x_{\min} = |O_{\min}| \tag{13}$$

We consider $o \notin O_{\min}$; in this case, either $\pi_o(r) < \alpha$ or $\pi_o(\bar{r}) \geq \alpha$. If $\pi_o(r) < \alpha$, however, it must be $\pi_o(\bar{r}) = 1 \geq \alpha$ since π_o is normal. Therefore, for each $o \notin O_{\min}$, $\pi_o(\bar{r}) \geq \alpha$, thus

$$\pi_{O_{\min}}(r) = \min\left\{\min_{o\in O_{\min}} \pi_o(r), \min_{o\notin O_{\min}} \pi_o(\bar{r})\right\} \geq \alpha$$

i.e. $x_{\min} \in [\pi_N]_\alpha$. Also, for $x_{\min} > 0$, if we remove a non-empty subset $O' \subseteq O_{\min}$ from $O_{\min}$, then:

$$\min_{o\in O_{\min}\backslash O'} \pi_o(r) \geq \alpha$$

but

$$\min_{o\notin O_{\min}\backslash O'} \pi_o(\bar{r}) = \min\left\{\min_{o\notin O_{\min}} \pi_o(\bar{r}), \min_{o\in O'} \pi_o(\bar{r})\right\} < \alpha$$

therefore:

$$\pi_{O_{\min}\backslash O'}(r) < \alpha$$

thus $|O_{\min}\backslash O'| = x_{\min} - k \notin [\pi_N]_\alpha$ for some $k > 0$. Therefore $x_{\min} = \min[\pi_N]_\alpha$.

We now define:

$$O_{\max} = \{o \in O | \pi_o(r) \geq \alpha\} \tag{14}$$

and

$$x_{\max} = |O_{\max}| \tag{15}$$

Of course $O_{\max} \supseteq O_{\min}$ therefore $x_{\max} \geq x_{\min}$. Also, if $o \notin O_{\max}$ then $o \notin O_{\min}$ therefore $\pi_o(\bar{r}) \geq \alpha$. As a consequence

$$\pi_{O_{\max}}(r) = \min\left\{\min_{o \in O_{\max}} \pi_o(r), \min_{o \notin O_{\max}} \pi_o(\bar{r})\right\} \geq \alpha$$

therefore $x_{\max} \in [\pi_N]_\alpha$. For $x_{\max} < m$, if we add a non-empty subset O' to $O_{\max}$ (such that $O' \cap O_{\max} = \emptyset$), then, by definition

$$\min_{o \in O_{\max} \cup O'} \pi_o(r) = \min\left\{\min_{o \in O_{\max}} \pi_o(r), \min_{o \in O'} \pi_o(r)\right\} < \alpha$$

since, by definition, if $o \notin O_{\max}$ then $\pi_o(r) < \alpha$. Therefore $x_{\max} + k = |O_{\max} \cup O'| \notin [\pi_N]_\alpha$ for $k > 0$, i.e. $x_{\max} = \max[\pi_N]_\alpha$.

Let $x \in \mathbb{N}$ such that $x_{\min} \leq x \leq x_{\max}$. Since $O_{\min} \subseteq O_{\max}$ we can identify a subset $O_k \subseteq O_{\max}$ such that $k = |O_k| = x - x_{\min}$ and $O_k \cap O_{\min} = \emptyset$. Let $O_x = O_{\min} \cup O_k$. Then, by definition:

$$\forall o \in O_x : \pi_o(r) \geq \alpha$$

because $O_x \subseteq O_{\max}$ and, since $O_k \cap O_{\min} = \emptyset$,

$$\forall o \notin O_x : o \notin O_{\min}$$

therefore:

$$\forall o \notin O_x : \pi_o(\bar{r}) \geq \alpha$$

As a consequence:

$$\pi_{O_x}(r) = \min\left\{ \min_{o \in O_x} \pi_o(r), \min_{o \notin O_x} \pi_o(\bar{r}) \right\} \geq \alpha$$

therefore, $x = x_{\min} + k = |O_{\min} \cup O_k| \in [\pi_N]_\alpha$, i.e. $\forall x \in \mathbb{N} : x_{\min} \leq x \leq x_{\max} \rightarrow x \in [\pi_N]_\alpha$.

In summary any α-cut of π_N is an interval in the domain of natural numbers, i.e.,

$$[\pi_N]_\alpha = [x_{\min}, x_{\max}]$$

Since an interval is a convex set, then, by definition, π_N is a convex fuzzy set. Since it also normal, then it is a fuzzy interval in the domain of natural numbers.

4. Algorithms for granular counting

To determine the granular count of a collection of observations, in principle we could directly apply its definition as in (7). However, this procedure would require the computation of all subsets O_x for any x from 0 to m. (m is the number of observations.) Since there are $\binom{m}{x}$ possible sets O_x, the total number of sets to be computed is

$$\sum_{x=0}^{m} \binom{m}{x} = 2^m$$

thus, the procedure has exponential time complexity, being intractable for real-world applications. Fortunately, there exists a polynomial algorithm to perform the computation, as described in Sec. 4.1. This algorithm computes the granular count with quadratic time complexity. Furthermore, a more efficient algorithm (with linear time complexity) is available if an approximate solution suffices. The approximate counting algorithm is described in Sec. 4.2.

4.1. Exact granular counting algorithm

Let

$$A = \left\{\pi | \pi = \pi_{o_j}(r_i)\, j = 1, 2, \ldots m,\, i = 1, 2, \ldots, n,\, \pi \neq 0\right\} \tag{16}$$

be the set of all distinct non-zero possibility degrees. It can be easily observed that $|A| \sim \mathcal{O}(nm)$.

For each $r \in R$, we first compute $\pi_o(\overline{r})$ for all $o \in O$. This procedure requires $\mathcal{O}(nm)$ comparisons and $\mathcal{O}(m)$ space. Let $\mathbf{n}_r \in [0,1]^{m+1}$ be an array of $m+1$ possibility degrees. Initially $\mathbf{n}_r \leftarrow \mathbf{0}$. Then, for each $\alpha \in A$, it is possible to compute the values $x_{\min}$ and $x_{\max}$ as in (13) and (15). In order to compute $x_{\min}$, it is necessary to compute $O_{\min}$ as in (12), which requires to scan all the values of $\pi_o(r)$ and the values of $\pi_o(\overline{r})$ (by varying $o \in O$). Both time and space complexity of this scan is $\mathcal{O}(m)$. The same procedure (thus, with the same complexity) is required to compute $x_{\max}$.

Since, for each $x \in [x_{\min}, x_{\max}]$, by construction $\pi_N(r) \geq \alpha$, then we can update the array $\mathbf{n}_r$ as

$$\mathbf{n}_r[x] \leftarrow \max\{\mathbf{n}_r[x], \alpha\} \tag{17}$$

requiring $\mathcal{O}(m)$ updates. At the end of the procedure, the array $\mathbf{n}_r$ will represent the granular count for referent r. The overall time complexity is $\mathcal{O}\left(nm^2\right)$, while the space complexity is $\mathcal{O}(nm)$. If the granular count must be performed for all the referents, the time complexity raises to $\mathcal{O}\left(n^2m^2\right)$.

Algorithm 1 is a pseudo-code of the exact counting algorithm. The referents can be represented as a matrix R, where each element r_{jl} corresponds to $\pi_{o_j}(r_l)$, i.e. the possibility degree that the j-th observation refers to the l-th referent. The algorithm EXACTGRANULARCOUNT requires the matrix R and the index

Algorithm 1: ExactGranularCount

Data: R, i
$R = [r_{jl}] = \left[\pi_{o_j}(r_l)\right]$, $l = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$
i is the index of the referent for which observations must be counted
Result: $\mathbf{n} = [n_x] \in [0, 1]$ for $x = 0, 1, \ldots, m$
1 $A \leftarrow \{r_{jl} : r_{jl} \neq 0,\ l = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m\}$ (16);
2 $\mathbf{n} \leftarrow$ granularCount(R, i, A);

i of the referent to be counted; it generates the set A as in (16) and calls granularCount (Algorithm 2).

The granularCount algorithm operates by computing $\bar{r}$ as in (8) (line 2) and initializing an array $\mathbf{n}$ of possibility degrees (line 3) which is eventually returned by the algorithm. Then, for each $\alpha \in A$, it computes the extremes of the interval $[x_{\min}, x_{\max}]$ (lines 7–11). Finally, all the values of $\mathbf{n}$ within the interval are updated according to (17) (lines 12–13).

Algorithm 2: granularCount

Data: R, i, A
$R = [r_{jl}] = \left[\pi_{o_j}(r_l)\right]$, $l = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$
i is the index of the referent for which observations must be counted
A is a set of α values
Result: $\mathbf{n} = [n_x] \in [0, 1]$ for $x = 0, 1, \ldots, m$
1 $\mathbf{r}_i \leftarrow [r_{ji}]_{j=1,2,\ldots,m}$ is the i-th column of R;
2 $\bar{\mathbf{r}}_i \leftarrow [\bar{r}_{ji}]_{j=1,2,\ldots,m}$ where $\bar{r}_{ji} = \max_{l \neq i} r_{jl}$ (8);
3 $\mathbf{n} \leftarrow [0, 0, \ldots, 0]$ ($m + 1$ times);
4 **for** $\alpha \in A$ **do**
5 $x_{\min} \leftarrow 0$;
6 $x_{\max} \leftarrow 0$;
7 **for** $k = 1, 2, \ldots, m$ **do**
8 **if** $r_{ki} \geq \alpha$ **then**
9 $x_{\max} \leftarrow x_{\max} + 1$;
10 **if** $\bar{r}_{ki} < \alpha$ **then**
11 $x_{\min} \leftarrow x_{\min} + 1$;
12 **for** $x \in x_{\min}, \ldots, x_{\max}$ **do**
13 $n_x \leftarrow \max\{n_x, \alpha\}$ (17);

Algorithm 3: APPROXIMATEGRANULARCOUNTING

Data: R, i
$R = [r_{jl}] = \left[\pi_{o_j}(r_l)\right]$, $l = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$
i is the index of the referent for which observations must be counted
Result: $\mathbf{n} = [n_x] \in [0, 1]$ for $x = 0, 1, \ldots, m$
1 $\varepsilon \leftarrow 10^{-12}$;
2 $A \leftarrow \{\varepsilon + k \cdot \frac{1-\varepsilon}{n_\alpha - 1} : k = 0, 1, \ldots, n_\alpha - 1\}$ (18);
3 $\mathbf{n} \leftarrow$GRANULARCOUNT(R, i, A);

4.2. Approximate granular counting algorithm

The computation of exact count requires a number of steps that is proportional to the cardinality of the set A, which is $\mathcal{O}(nm)$. (This can be easily observed from Algorithm 1.) As a result, the computational time complexity of the exact counting algorithm is quadratic with the number of observations. In real-world scenarios, it is expected that the number of observations can be very large (on the other hand, the number of referents is expected to be smaller), therefore an algorithm with quadratic complexity may be inefficient.

It is possible to cut down the time complexity by resorting to an approximate counting algorithm. This variants simply does not compute the set A as in (16), rather a number $n_\alpha > 1$ is required and the set $\widetilde{A}$ is defined by equidistant values as follows:

$$\widetilde{A} = \left\{ \varepsilon + k \cdot \frac{1-\varepsilon}{n_\alpha - 1} \middle| k = 0, 1, \ldots, n_\alpha - 1 \right\} \tag{18}$$

i.e. $\widetilde{A} = \left\{ \varepsilon, \varepsilon + \frac{1-\varepsilon}{n_\alpha - 1}, \varepsilon + 2 \cdot \frac{1-\varepsilon}{n_\alpha - 1}, \ldots, 1 \right\}$. We use $\varepsilon > 0$ as the smallest value of α in order to avoid the degenerate case $\alpha = 0$. A very small value of ε must be used. In Algorithm 3 the pseudo-code of the approximate counting method is presented, where $\varepsilon = 10^{-12}$.

The time complexity of the approximate counting algorithm is $\mathcal{O}(m)$, i.e. it is linear with the number of observations. (In the case that the approximate counting algorithm is applied to all the referents, the time complexity is

increased to $\mathcal{O}(mn)$.) This makes the approximate granular counting method very efficient when a large number of observations is available. However, it requires an empirical determination of the parameter n_α.

5. Experimental results

The experimentation has two objectives. The first objective is to compare the approximate and exact counting algorithms in terms of similarity of the resulting fuzzy sets. To this aim, we adopt a synthetic data table and a collection of similarity measures. The second objective is to demonstrate that the counting method is useful in a real-world scenario. To this end, we apply the counting algorithms to public Bioinformatics datasets concerning sequenced RNA.

5.1. Comparison of exact and approximate granular counting algorithms

The objective of this experiment is to evaluate how much dissimilar the fuzzy sets are resulting from the application of the exact and approximate granular counting algorithms on a synthetic data table. (The use of synthetic data is motivated by the possibility of manually checking the correctness of the results.)

The data table we used is reported in table 1. It is defined by ten observations and three referents, with a total of 11 distinct possibility values (excluding 0). The choice of the values is motivated by the need of having a non-trivial, yet not too large set of possibility values that are not equally spaced in the interval $[0, 1]$. In this way, it is impossible that a run of the approximate algorithm generates α-cuts with boundaries coinciding with the possibility shown values in the table. This makes the experiment closer to a real-world scenario.

In the literature, there is not a general agreement on the definition of similarity between fuzzy sets; therefore, in order to avoid to bias the results to a specific similarity measure, we measured the similarity of the results according to several similarity measures. In table 2 the collection of similarity measures,

Table 1: Synthetic data table used for comparing approximate vs. exact counting algorithms

	r_1	r_2	r_3
o_1	1	.3	.54
o_2	.8	1	.6
o_3	1	0	0
o_4	.86	.91	1
o_5	1	0	0
o_6	.5	1	.64
o_7	1	.8	1
o_8	.2	.5	1
o_9	1	0	0
o_{10}	.6	1	.78

Table 2: Similarity measures used in the first experiment.

Similarity measure	**Short description**
$J = \frac{\sum_{i=1}^{n} \min\{a_i, b_i\}}{\sum_{i=1}^{n} \max\{a_i, b_i\}}$	set-based similarity
$L = 1 - \max_i(\lvert a_i - b_i \rvert)$	maximum difference of membership degrees
$S = 1 - \frac{\sum_i \lvert a_i - b_i \rvert}{\sum_i (a_i + b_i)}$	sum of differences of membership degrees
$W = 1 - \frac{\sum_{i=1}^{n} \lvert a_i - b_i \rvert}{n}$	distance-based similarity
$P = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\max\left\{\sum_{i=1}^{n} a_i^2, \sum_{i=1}^{n} b_i^2\right\}}$	matching-based similarity
note: $A = \sum_{i=1}^{n} a_i/x_i$, $B = \sum_{i=1}^{n} b_i/x_i$ in Zadeh's notation of fuzzy sets.	

used in the experiment, is reported. For further details, the interested reader is referred to the specialized literature [29, 30].

We applied the exact and approximate granular counting algorithms to all the referents in table 1. The approximate counting algorithm requires the number n_α of α-cuts to be computed. Given the dataset at hand, we run the approximate counting algorithm with three values of n_α, namely 2, 5, 10. With $n_\alpha = 2$ it is possible to approximate the resulting fuzzy set with a trapezoidal shape that can be conveniently used in many software tools for fuzzy inference. On the other hand, the approximate results may be too dissimilar from the exact solution, which can be only approximated with a larger number of α-cuts.

In figs. 1, 2 and 3, the results of counting are depicted. According to the exact counting algorithm, it is impossible that r_1 is referred by less than 3

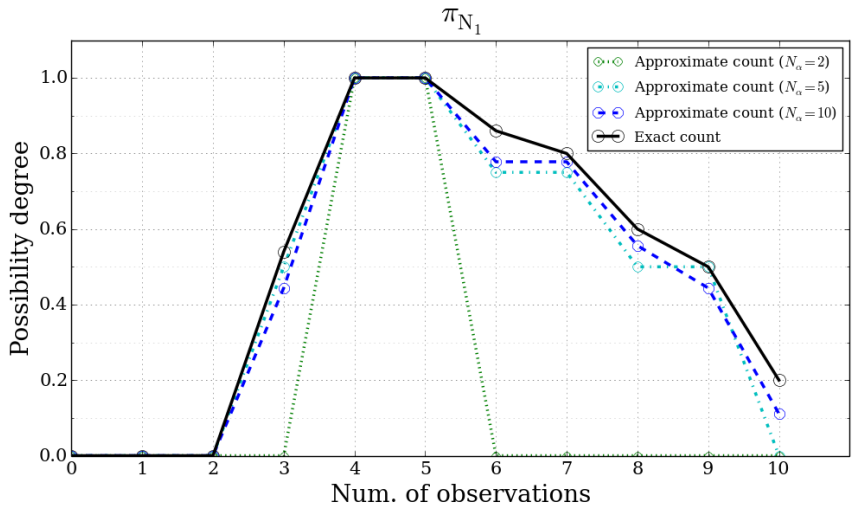

Figure 1: Granular counting (exact and approximate) of referent r_1

observations (indeed, o_3, o_5 and o_9 can only refer to r_1 according to table 1), as depicted in fig. 1; however, since o_1 is more likely to refer to r_1 than to r_2 or r_3, then the possibility degree $\pi_{N_1}(x)$ on the actual number of observations counting for r_1 is maximal for $x = 4$, while it is in the middle for $x = 3$ (more precisely, $\pi_{N_1}(3) = 0.54$). This behavior is well represented by the approximate counts for $n_\alpha = 5, 10$ but it is not captured for $n_\alpha = 2$. In the latter case, the approximate counting returns full possibility that the number of observations for r_1 is four, but null possibility for three (or less) observations. Overall, approximate counts with $n_\alpha = 5, 10$ well follow the shape of the exact counting, while approximate counting with $n_\alpha = 2$ promotes efficiency over accuracy. This behavior is confirmed in figs. 2 and 3.

In order to quantify the accuracy of the approximate counting, we run the algorithm for an increasing number of α-cuts, ranging from 2 to 100. For each run, we evaluated the similarity of the approximate count with the exact count according to the similarity measures reported in table 2. The results are depicted in figs. 4, 5 and 6.

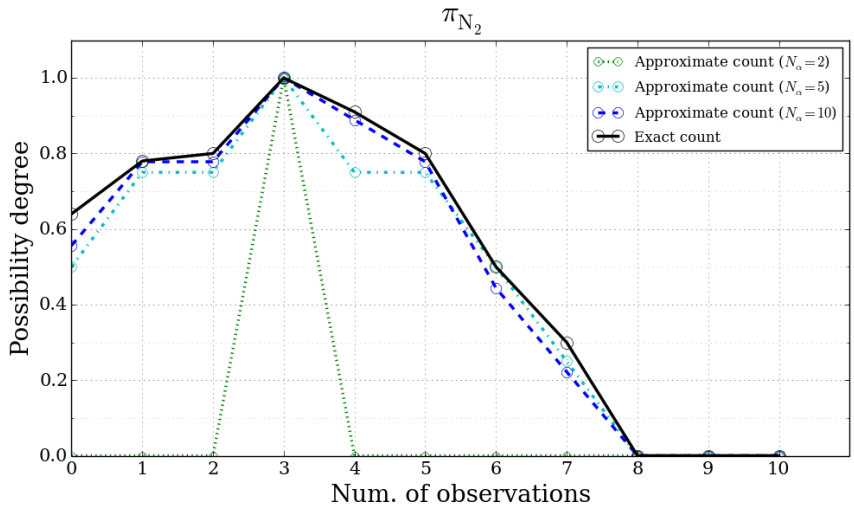

Figure 2: Granular counting (exact and approximate) of referent r_2

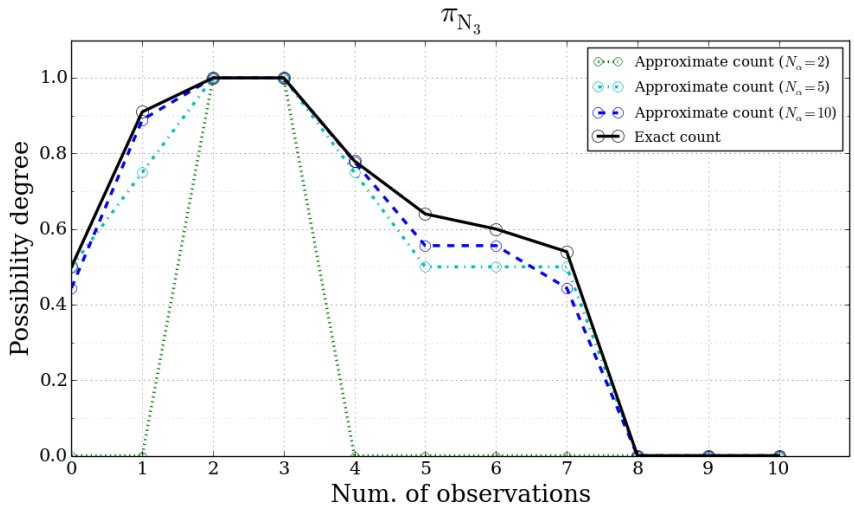

Figure 3: Granular counting (exact and approximate) of referent r_3

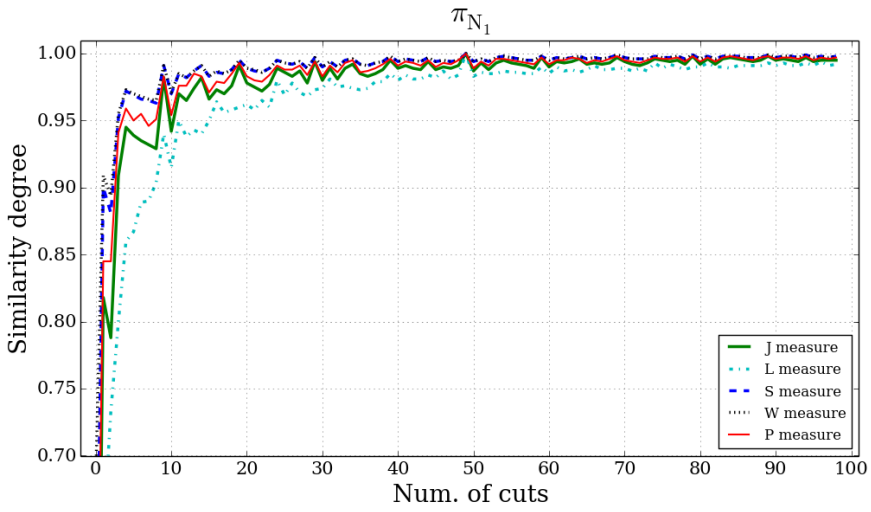

Figure 4: Similarity values of approximate counting for r_1

We observe that all the similarity measures show almost the same behavior, with the exception of L similarity (perhaps due to the lack of a summation operation). The trends are not monotonic but oscillating, with irregular amplitudes that tend to damp for large numbers of α-cuts. This behavior has a simple explanation: since the α-levels generated by the approximate counting algorithm are equidistant in $[0, 1]$, it is unlikely that they coincide with the possibility degrees in the dataset. This imperfect matching generates oscillations. As the number of α-cuts increases, the corresponding α-levels become closer to the values present in the dataset, thus determining the attenuation of oscillations. For a theoretically infinite number of α-cuts, it is expected that all the possibility degrees in the dataset correspond to some α-level, thus the approximate counting asymptotically becomes an exact counting.

5.2. Real world example: counting multi-reads in NGS

One of the main promoters of the thriving development of Bioinformatics is the management and analysis of large amounts of data. Genomes, genes,

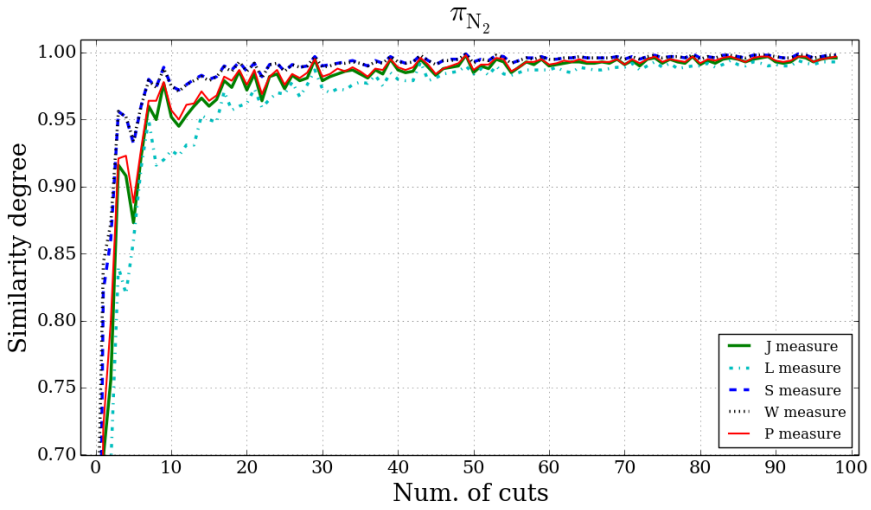

Figure 5: Similarity values of approximate counting for r_2

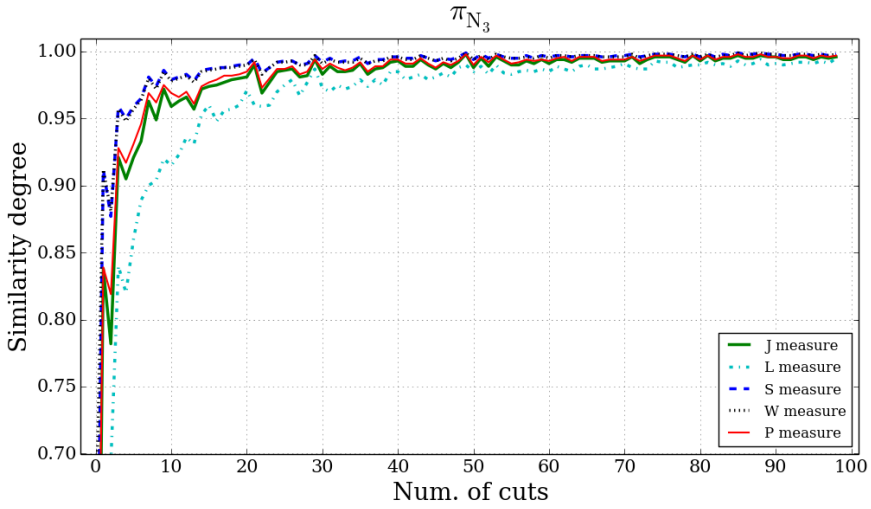

Figure 6: Similarity values of approximate counting for r_3

transcripts, proteins, pathways, ad many other experimental data fill up the biological databases and demand efficient and fast algorithms to be analyzed. The advent of Next Generation Sequencing (NGS) has hastened the production of molecular data by decoding the DNA and RNA sequences contained in cell samples. RNA-Seq is a NGS protocol that allows to examine the RNA content of a cell and to transcribe it in a text file. The output of the process is a list of RNA sequences decoded in strings (called "reads"). The comparison of different samples aimed at extracting the differences in the number of copies of each gene (the gene expression) is known as Differential Expression Analysis (DEA) [31].

When RNA-Seq output is mapped against a reference database, a high percentage of reads (even more than 30%) map to more than one gene. This reads are called "multireads". They occur mainly because genes share repeated portions and because reads are usually shorter than original RNA transcripts. If a read maps equally to two or more genes that have an identical portion in common, there is no way to prove that the read originates effectively from one of the two genes. Multireads can be a source of uncertainty in the quantification of gene expression, and this uncertainty should be represented and managed, instead of trying to remove it, because errors in read count estimations can lead to false positives in the results [32]. NGS machines provide a "quality" index that, according to biologists, is a biologically plausible estimate of the possibility that a read can be associated to a gene. However, a high quality index does not mean certainty in association: two or more genes can be candidate for mapping a read because they can be mapped with similar high quality.

The use of probability theory for modeling this form of uncertainty is not advisable for a couple of reasons: (i) probability could be estimated by normalizing the quality of a mapping over the sum of all quality indexes of a read mapping to all genes. However, in this way, even if the mapping quality of a

read to a gene is very high, the probability could be low if several genes have high quality indexes: this could still be a correct choice if we can verify with a statistical approach if the quality index determines the relative frequencies of the associations, which in turn estimate probability, but unfortunately these repeated experiments are not available in DEA. (ii) By using possibility, which is a less stringenttheory than probability for handling uncertainty, it is possible to devise an approximate counting algorithm that runs in linear time on the number of reads, which can be hundreds of thousands. To the authors' knowledge, there is not a corresponding algorithm using probabilities.

It is possible to apply the counting method presented in this paper to achieve a granular representation of the read counts. This granular representation can be eventually used to perform differential analysis by taking into account the uncertainty in counting, instead of artificially removing it. (We already applied a preliminary version of granular counting in this scenario [33], but the method was defined on empirical bases and produced trapezoidal fuzzy sets only.)

We used the public dataset SRP014005 downloaded from NCBI-SRA archive[4], which contains a case-control study of the Asthma disease, performed through 454 Roche sequencing of human endobronchial biopsies. The 55,579 reads were mapped on 14,802 genes contained in the Vega transcript database [34] by using BLAST (with 97% of identity required). A total of 7,725 reads (i.e. 16% of the entire dataset) resulted as multireads. The working assumption we used in our experiment is that the higher is the mapping accuracy, the higher is the possibility that a read actually belongs to a gene. Based on this assumption, the possibility degree that the a read actually maps on a gene is given by the product of "identity" and "coverage" (two quality indexes returned by BLAST),

[4] `ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP014/SRP014005`

eventually normalized so that for each read at least one gene has full possibility of mapping. According to this setting, each read plays the role of observation for our counting method, while each gene acts as a referent.

For each gene, it is possible to compute the granular counting of the reads. As an illustrative example, we consider gene OTTHUMG00000165694|RIC3, which is mapped by 59 reads also mapping to other 391 genes. In fig. 7 the results of granular counting are reported, by considering exact granular counting as well as approximate granular counting with 2, 3, 5 and 10 α-cuts. What is immediate to observe is the high uncertainty showed by the granular representation of counts: any analysis method that does not take into account such uncertainty fails to provide veridical results.

We also observe when too few α-cuts are used (2-3), the approximate counting carries out results that are very distant from the exact counting. However, few more α-cuts (10) are enough to provide an accurate result that can be effectively used in subsequent operations. This result is confirmed by computing the Jaccard similarity measure (J as in table 2) between exact and approximate counts for a sample of genes. As can be observed from table 3, with 10 α-cuts the similarity values are always higher than 0.9. Of course, a greater number of α-cuts give better results in terms of similarity, but the price to pay is in terms of computing time[5].

6. Conclusion

In this paper we have presented a method for counting uncertain data, when uncertainty is modeled through possibility distributions. We also devised two algorithms for granular counting: an exact algorithm with quadratic time com-

[5]All the experiments run for this dataset required less than five seconds on a standard office computer. The time is referred to running the granular counting algorithm. The time required for data preparation has not been taken into account.

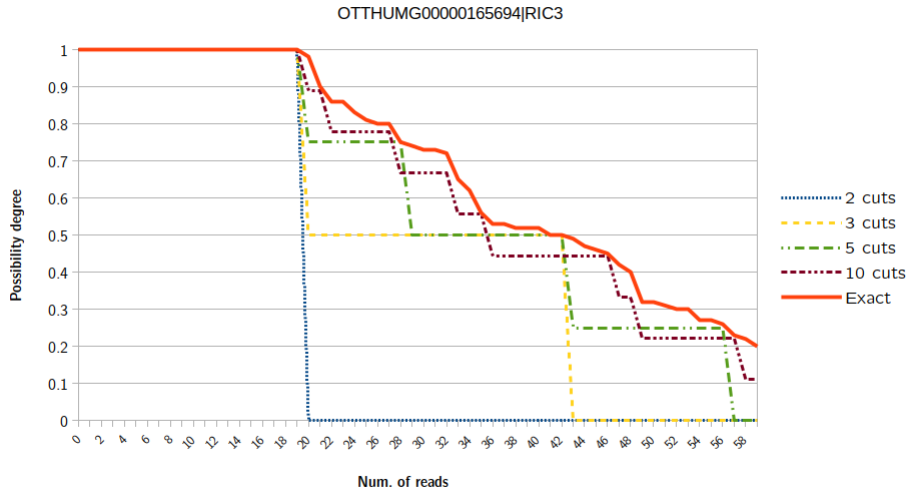

Figure 7: Granular counting of reads mapping to a sample gene

Table 3: Jaccard (J) similarity measures between approximate and exact granular counting for a selection of genes

Gene	2-cuts	3-cuts	5-cuts	10-cuts
OTTHUMG00000168110|C2CD3	0.76	0.89	0.96	0.98
OTTHUMG00000131426|SLC28A2	0.36	0.59	0.83	0.93
OTTHUMG00000133170|BRD7	0.62	0.77	0.89	0.96
OTTHUMG00000153289|RGPD8	0.50	0.76	0.88	0.95
OTTHUMG00000129407|SULT1B1	0.27	0.64	0.83	0.94
OTTHUMG00000021744|OPHN1	0.31	0.66	0.84	0.94
OTTHUMG00000161751|RP11-703G6.1	0.65	0.79	0.91	0.97
OTTHUMG00000020565|RP11-442A13.1	0.02	0.59	0.80	0.92
OTTHUMG00000165694|RIC3	0.48	0.76	0.89	0.94

plexity and an approximate algorithm with linear time complexity, which have been empirically compared. The illustrative case in Bioinformatics shows that granular counting may provide useful information by propagating uncertainty in the primary source of data (reads in NGS) towards the subsequent stages of analysis, which rely on read counting to estimate the expression of genes in cells. We observed that the resulting granular counts show a low specificity as a result of the high uncertainty in the reads. It is interesting to note that common practice in Bioinformatics concerns counting just by disregarding uncertainty and returning a single value [33]. The use of granular counting may shed light on more comprehensive analyses and more significant results.

The counting algorithms that have been devised are computationally efficient. This result has been possible thanks to the use of Possibility Theory for modeling uncertainty. To the authors' knowledge, there are not linear-time algorithms for estimating counts in the case that uncertainty is modeled with probability distributions. Future research will stress on the efficiency of counting algorithms by applying them in real-world scenarios involving big data, such as multireads counting in realistic scenarios with a number of reads of order 10^8 in magnitude (and constantly increasing thanks to technological advancements), and a number of genes of order 10^5 in magnitude. Finally, future research will be devoted to the study of the theoretical connections of the proposed method with other formal frameworks based on Possibility Theory, including twofold fuzzy sets [35].

Acknowledgments

This research was partially supported by the Canada-Italy Innovation Award 2016 granted by the Government of Canada. C.M. thanks the student Mrs. Annarita Fierro for her support in performing experiments.

References

References

[1] R. Kitchin, The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences, SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2014. doi:10.4135/9781473909472.

[2] C. C. Aggarwal, P. S. Yu, A Survey of Uncertain Data Algorithms and Applications, IEEE Transactions on Knowledge and Data Engineering 21 (5) (2009) 609–623. doi:10.1109/TKDE.2008.190.

[3] N. Boukhelifa, M.-E. Perrin, S. Huron, J. Eagan, How Data Workers Cope with Uncertainty: A Task Characterisation Study, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, 2017, pp. 3645–3656. doi:10.1145/3025453.3025738.

[4] P. Kakar, A. Y.-S. Chia, If You Can't Beat Them, Join Them, in: Proceedings of the 23rd ACM international conference on Multimedia - MM '15, ACM, ACM Press, New York, New York, USA, 2015, pp. 571–580. doi:10.1145/2733373.2806231.

[5] A. Ghosh, N. Manwani, P. S. Sastry, Making risk minimization tolerant to label noise, Neurocomputing 160 (2015) 93–107. doi:10.1016/j.neucom.2014.09.081.

[6] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE transactions on neural networks and learning systems 25 (5) (2014) 845–869. doi:10.1109/TNNLS.2013.2292894.

[7] X. Geng, Label Distribution Learning, IEEE Transactions on Knowledge and Data Engineering 28 (7) (2016) 1734–1748. arXiv:1702.06086, doi:

10.1109/TKDE.2016.2545658.

URL http://arxiv.org/abs/1408.6027

[8] P. K. Agarwal, S.-W. Cheng, Y. Tao, K. Yi, Indexing uncertain data, in: Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, 2009, pp. 137–146. doi:10.1145/1559795.1559816.

[9] L. Sun, R. Cheng, D. W. Cheung, J. Cheng, Mining uncertain data with probabilistic guarantees, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10, ACM, ACM Press, New York, New York, USA, 2010, p. 273. doi:10.1145/1835804.1835841.

[10] M. Hua, J. Pei, W. Zhang, X. Lin, Ranking queries on uncertain data: a probabilistic threshold approach, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, 2008, pp. 673–686. doi:10.1145/1376616.1376685.

[11] T. Cour, B. Sapp, C. Jordan, B. Taskar, Learning from ambiguously labeled images, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 919–926. doi:10.1109/CVPR.2009.5206667.
URL https://ieeexplore.ieee.org/document/5206667/

[12] T. Y. Lin, N. Cercone, Rough Sets and Data Mining, Springer US, Boston, MA, 1996. doi:10.1007/978-1-4613-1461-5.

[13] P. Vannoorenberghe, Reasoning with unlabeled samples and belief functions, in: The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ '03., Vol. 2, IEEE, 2003, pp. 814–818. doi:10.1109/FUZZ.2003.1206534.
URL http://ieeexplore.ieee.org/document/1206534/

[14] D. Dubois, H. Prade, Possibility Theory, in: Computational Complexity, Springer New York, New York, NY, 2012, pp. 2240–2252. doi:10.1007/978-1-4614-1800-9_139.
URL http://link.springer.com/10.1007/978-1-4614-1800-9_139

[15] D. Dubois, Possibility theory and statistical reasoning, Computational Statistics & Data Analysis 51 (2006) 47–69. doi:10.1016/j.csda.2006.04.015.

[16] F. Delmotte, Detection of defective sources in the setting of possibility theory, Fuzzy Sets and Systems 158 (5) (2007) 555–571. doi:10.1016/j.fss.2006.10.027.

[17] S. Benferhat, K. Tabia, Inference in possibilistic network classifiers under uncertain observations, Annals of Mathematics and Artificial Intelligence 64 (2) (2012) 269–309. doi:10.1007/s10472-012-9290-1.

[18] J. Hulsmann, A. Buschermohle, W. Brockmann, Incorporating Dynamic Uncertainties into a Fuzzy Classifier, in: Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011), Atlantis Press, Paris, France, 2011, pp. 388–395. doi:10.2991/eusflat.2011.4.

[19] M. Bounhas, M. G. Hamed, H. Prade, M. Serrurier, K. Mellouli, Naive possibilistic classifiers for imprecise or uncertain numerical data, Fuzzy Sets and Systems 239 (Supplement C) (2014) 137–156. doi:10.1016/j.fss.2013.07.012.

[20] L. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1) (1978) 3–28. doi:10.1016/0165-0114(78)90029-5.
URL http://linkinghub.elsevier.com/retrieve/pii/0165011478900295

[21] H. Bandemer, W. Näther, Fuzzy data analysis, Vol. 20, Springer Science & Business Media, 2012. doi:10.1007/978-94-010-0646-0_13.

[22] D. Dubois, H. Prade, Fuzzy cardinality and the modeling of imprecise quantification, Fuzzy sets and Systems 16 (3) (1985) 199–230. doi:10.1016/0165-0114(85)90025-9.

[23] B. Kosko, Counting with Fuzzy Sets, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8 (4) (1986) 556–557. doi:10.1109/TPAMI.1986.4767822.

[24] D. Ralescu, Cardinality, quantifiers, and the aggregation of fuzzy criteria, Fuzzy sets and systems 69 (3) (1995) 355–365. doi:10.1016/0165-0114(94)00177-9.

[25] L. A. Zadeh, Possibility Theory and Soft Data Analysis, in: Fuzzy sets, fuzzy logic, and fuzzy systems, 1996, pp. 481–541. doi:10.1142/9789814261302_0025.

[26] R. Wille, Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies, in: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis: Foundations and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 1–33. doi:10.1007/11528784_1.
URL https://doi.org/10.1007/11528784_1

[27] D. Dubois, H. Prade, Possibility theory and formal concept analysis: Characterizing independent sub-contexts, Fuzzy Sets and Systems 196 (2012) 4–16. doi:10.1016/J.FSS.2011.02.008.

[28] D. Dubois, H. Prade, Bridging gaps between several forms of granular computing, Granular Computing 1 (2) (2016) 115–126. doi:10.1007/

s41066-015-0008-8.

URL http://link.springer.com/10.1007/s41066-015-0008-8

[29] S.-M. Chen, M.-S. Yeh, P.-Y. Hsiao, A Comparison of Similarity Measures of Fuzzy Values, Fuzzy Sets Syst. 72 (1) (1995) 79–89. doi:10.1016/0165-0114(94)00284-E.

[30] C. P. Pappis, N. I. Karacapilidis, A comparative assessment of measures of similarity of fuzzy values, Fuzzy Sets and Systems 56 (2) (1993) 171–174. doi:10.1016/0165-0114(93)90141-4.

[31] A. Consiglio, C. Mencar, G. Grillo, S. Liuni, Managing NGS Differential Expression Uncertainty with Fuzzy Sets, in: C. Angelini, S. Rovetta, P. M. V. Rancoita (Eds.), Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2015 (Revised Selected Papers), Vol. 9874 of Lecture Notes in Bioinformatics, Springer, Naples, Italy, 2016, pp. 42–53. doi:10.1007/978-3-319-44332-4_4.

URL http://link.springer.com/10.1007/978-3-319-44332-4_4

[32] A. Consiglio, MultiDEA: a fuzzy method for RNA-Seq Differential Expression Analysis in presence of Multireads, Ph.d. thesis, University of Bari "A. Moro" (2016).

[33] A. Consiglio, C. Mencar, G. Grillo, F. Marzano, M. F. Caratozzolo, S. Liuni, A fuzzy method for RNA-Seq differential expression analysis in presence of multireads, BMC Bioinformatics 17 (S12:345) (2016) 167–182. doi:10.1186/s12859-016-1195-2.

URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1195-2

[34] L. G. Wilming, J. G. R. Gilbert, K. Howe, S. Trevanion, T. Hubbard, J. L.

Harrow, The vertebrate genome annotation (Vega) database, Nucleic acids research 36 (suppl. 1) (2007) D753–D760. doi:10.1093/nar/gkm987.

[35] D. Dubois, H. Prade, Twofold fuzzy sets and rough sets-Some issues in knowledge representation, Fuzzy Sets and Systems 23 (1) (1987) 3–18. doi:10.1016/0165-0114(87)90096-0.
URL https://www.sciencedirect.com/science/article/pii/0165011487900960