MethodsX 6 (2019) 2455–2459

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

GrCount: Counting method for uncertain data

Corrado Mencar[a,*], Witold 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

A B S T R A C T

We report a method for counting uncertain data, i.e. observations that cannot be precisely associated to referents. We model data uncertainty through Possibility Theory and we develop the counting method so as to take into account the possibility distributions attached to data. The result is a fuzzy interval on the domain of natural numbers, which can be obtained by two variants of the method: exact counting provides the true fuzzy interval in quadratic time complexity, while approximate counting carries out an estimate of the fuzzy interval in linear time. We give a step-by-step description of the method so that it can be replicated in any programming environment. We also provide a Python implementation and a use case in Bioinformatics. The method usage is the following:

- The uncertain data are represented in form of matrix, one row for each observation. Each row is a possibility distribution;
- The method variant must be selected. In the case of the approximate variant, the number of α-values of the resulting fuzzy interval must be provided;
- For each referent, a fuzzy interval is determined and carried out by the method.

A R T I C L E I N F O
Method name: GrCount
Keywords: Granular computing, Counting, Possibility theory, Fuzzy intervals
Article history: Received 26 April 2019; Accepted 3 October 2019; Available online 17 October 2019

DOI of original article: http://dx.doi.org/10.1016/j.fss.2019.04.018
* Corresponding author.
E-mail address: corrado.mencar@uniba.it (C. Mencar).

https://doi.org/10.1016/j.mex.2019.10.001

Specification Table

Subject Area:	*Computer Science*
More specific subject area:	*Granular Computing*
Method name:	*GrCount*
Name and reference of original method:	*Corrado Mencar, Witold Pedrycz. Granular Counting of Uncertain Data. Fuzzy Sets and Systems,* https://doi.org/10.1016/j.fss.2019.04.018
Resource availability:	*The supplemental dataset ASTHMA_CTRL-read_gene_score-NORMALIZED.csv can be used to test the reproducibility of results.*

Data uncertainty

Data analysis requires the collection of possibly large amounts of data. In very general settings, data collection requires the observation of a phenomenon and the subsequent reference of such observation to one of possible referents. (A simple example is the reference of the height of a person to one of possible height values; a more complex example is the reference of a RNA fragment to a gene expressed in a cell.) In many cases, however, the reference of an observation to a referent can be uncertain, i.e. it is not possible to establish the referent of an observation in unequivocal way. We call such uncertainty as data uncertainty.

There are several reasons for uncertainty present in data [1,2]; reasonably, the more complex is the phenomenon under investigation, the more likely collected data are uncertain. A simple and common approach to deal with data uncertainty is to ignore uncertainty, but this strategy is dangerous as it may introduce bias in the subsequent processing stages, which is hard to recognize. Uncertainty can be actually 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.

Possibility Theory is a mathematical theory which is effective dealing with uncertainty due to incomplete information [3]. In essence, this theory enables the definition of a possibility distribution π over a set U of objects (or values); this possibility distribution represents the possibility that a variable X has a value x in U. In other words, the variable X has a value that is not known; what is known are the values that cannot be assigned to X and the values that may be assigned to X, although not necessarily. According to Possibility Theory, for each element x of U, $\pi(x) = 1$ if x is a possible value for X and $\pi(x) = 0$ if x cannot be a value of X. Most importantly, $\pi(x)$ can assume degrees from 0 to 1, i.e. the possibility can be graded. The degree of possibility is determined according to the problem at hand; for example, possibility can be connected to the similarity of x w.r.t. a prototype. From a possibility distribution, a possibility measure can be defined on the powerset of U such that, for any A subset of U, $\Pi(A)$ is the possibility that X has value in A, i.e. $\Pi(A) = \max\{\pi(x) \mid x \text{ in } A\}$. When more than one variables are involved, say X and Y with corresponding possibility distributions, then the possibility that X has value in A and Y has value in B is given by $\min\{\Pi_X(A), \Pi_Y(B)\}$ provided that X and Y are not interactive, i.e. the knowledge of the value of one variable does not affect the knowledge of the value of the other variable. Possibility Theory is indicated in the case of incomplete information, i.e. when a variable must take one value, but all that is known is a set of possible values. Graded possibility may come up by subjective evaluation (plausibility) or by feasibility assessment [4]. Possibility Theory is complementary to Probability Theory, which models random phenomena; in fact, the two theories can be combined giving rise to complex models of uncertainty [5].

A basic operation with data is counting, i.e. finding the number of data samples having a specific value. Counting is non-trivial when data are uncertain; in fact, uncertainty in data must propagate in counting, therefore results are granular rather than precise. In the companion paper [6] we report a formal analysis of granular counting and we prove that, when uncertainty is modeled through Possibility Theory, granular counting leads to fuzzy intervals in the domain of natural numbers. We also report two algorithms for granular counting: an exact algorithm which has quadratic complexity and an approximate algorithm with linear time complexity. In this paper we focus on the counting method so that it can be implemented and replicated in any programming environment.

Terminology and notation

We assume the existence of a collection of objects or referents $r_i, i = 1, 2, ..., n$, which are detected through observations $o_j.j = 1, 2, ..., m$. We assume that uncertainty is due to incomplete information coming from an observation, which impedes a unequivocal reference to one of the referents. For example, when observing a fragment of RNA as a result of the expression of a gene in a cell, in many cases it is not possible to refer unequivocally to the expressed gene because the same RNA fragment can be mapped to several genes with different degrees of possibility.

In essence, an observation is represented as a tuple $[\pi_{j1}, \pi_{j2}, ..., \pi_{jn}]$, where π_{ji} is the possibility degree that the observation o_j refers to referent r_i. Since we usually have many observations, we can collect of all them into a matrix

$$R = [\pi_{ji}]$$

with the constraint that for each row j there exists at least one index i such that $\pi_{ji} = 1$. (This is the normality axiom of a possibility distribution.) Following is an example of matrix collecting the possibility degrees of 5 observations and 3 referents:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 1 & .8 & .6 \\ 1 & 1 & 0 \\ .6 & .8 & 1 \\ .5 & 1 & 0 \end{bmatrix}$$

Notice that, according to Possibility Theory, $\pi_{ji} = 0$ means that it is *impossible* that observation o_j refers to referent r_i; on the other hand, $\pi_{ji} = 1$ means that it is *possible* (though not certain) that o_j refers to referent r_i. Intermediate values represent degrees of possibility, which may reflect a degree of feasibility or state of knowledge [7]. In the Supplementary material section we report a real-world example concerning the representation of uncertain data in the domain of Bioinformatics, which is also described in our companion paper.

Granular counting

The granular counting methods takes in input the matrix R and the column index i of the referent for which observations must be counted. It then returns a fuzzy interval on natural numbers representing the granular count. A fuzzy interval is a fuzzy set on the domain of natural number, i.e. a function

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

that is convex and normal. Informally speaking, this means that there exists an interval of natural numbers such that $N(x) = 1$ for all elements of this interval. Furthermore, the function N is monotonically increasing at the left of this interval and monotonically decreasing at its right.

The determination of the function N is carried out by computing its α-cuts. An α-cut of a fuzzy set is the set

$$[N]_\alpha = \{x : N(x) \geq \alpha\}$$

where $0 < \alpha \leq 1$. It is possible to prove that, by varying α, the fuzzy set N can be completely determined.

The first step is to determine the set of values of α that are required to define the corresponding α-cuts which, in turn, will be used to represent the fuzzy interval. There are two versions of granular counting: exact and approximate. Their main difference stands in the computation of the values of α that are used to determine the fuzzy set N: in exact counting, they are automatically determined from the matrix R, while in the approximate version a set of n_α equally spaced values of α are generated, where n_α is provided by the user. In particular:

- In the case of exact counting, the set A of α-values is given by $A = \{\pi_{ji} : i = 1, 2, \ldots n, j = 1, 2, \ldots m, \pi_{ji} \neq 0\}$
- In the case of approximate counting, the set A is defined as $A = \{\varepsilon + k \cdot \frac{1-\varepsilon}{n_\alpha - 1} : k = 0, 1, \ldots, n_\alpha - 1\}$

By continuing the previous example, we have:
(exact) $A = \{.5, .6, .8, 1\}$
(approximate) $A = \{\varepsilon, .505, 1\}$ with $\varepsilon = 10^{-12}$ and $n_\alpha = 3$

The construction of the fuzzy interval is carried out for each i-th referent as follows. From the matrix R two column vectors are obtained: the i -th column vector r_i with elements $r_{ji} = \pi_{ji}$ and the column vector $\bar{r}_i$ with elements $\bar{r}_{ji} = max_{k \neq i} \pi_{jk}$. For instance, if $i = 1$ and R is the matrix in the example, then:

$$r_1 = [1, 1, 1, .6, .5]^\top \quad \text{and} \quad \bar{r}_1 = [0, .8, 1, 1, 1]^\top$$

The fuzzy interval can be conveniently represented as an array of $m + 1$ membership degrees. We define the row vector v of size $1 \times (m + 1)$ which is initialized to zero, i.e. $v_j \leftarrow 0$ for $j = 0, 1, 2, \ldots, m$. (For the sake of simplicity, we start indexing v from 0.) The vector v is updated by cycling on the elements of A as follows.

For each $\alpha \in A$, the following sets are determined:

- $O_{min} = \{k : r_{ki} \geq \alpha \wedge \bar{r}_{ki} < \alpha\}$
- $O_{max} = \{k : r_{ki} \geq \alpha\}$

We are actually interested in the cardinality of the two sets, namely:

- $x_{min} = |O_{min}|$
- $x_{max} = |O_{max}|$

Then, the vector v is updated as follows: for each $x_{min} \leq j \leq x_{max} : v_j \leftarrow max\{v_j, \alpha\}$. The procedure is repeated until all values of A have been scanned. The returned function N is defined as:

$$N(x) = v_x \quad \text{if } 0 \leq x \leq m, \text{ otherwise} \quad N(x) = 0$$

It should be noticed that the number of elementary operations in each iteration of the procedure is $O(m)$ and the number of iterations is $O(|A|)$. In exact counting, $|A|$ is $O(nm)$ therefore the time complexity of the entire procedure is $O(nm^2)$, while in approximate counting $|A|$ is $O(1)$ therefore the overall time complexity is still $O(m)$.

Following is an example of step-by-step determination of the vector v in accordance to the example matrix for $i = 1$. We first assume exact counting.

- Initially, $v = [0, 0, 0, 0, 0, 0]$
- For $\alpha = .5$, $O_{min} = \{1\}$ therefore $x_{min} = 1$; $O_{max} = \{1, 2, 3, 4, 5\}$ therefore $x_{max} = 5$. The vector is updated to $v = [0, .5, .5, .5, .5, .5]$
- For $\alpha = .6$, $O_{min} = \{1\}$ therefore $x_{min} = 1$; $O_{max} = \{1, 2, 3, 4\}$ therefore $x_{max} = 4$. The vector is updated to $v = [0, .6, .6, .6, .6, .5]$
- For $\alpha = .8$, $O_{min} = \{1\}$ therefore $x_{min} = 1$; $O_{max} = \{1, 2, 3\}$ therefore $x_{max} = 3$. The vector is updated to $v = [0, .8, .8, .8, .6, .5]$
- For $\alpha = 1$, $O_{min} = \{1, 2\}$ therefore $x_{min} = 2$; $O_{max} = \{1, 2, 3\}$ therefore $x_{max} = 3$. The vector is updated to $v = [0, .8, 1, 1, .6, .5]$

It is noteworthy observing that the computation of v is irrespective of the order in which the values of A are considered.

In the case of approximate counting, the process is the same: just the α-values change:

- Initially, $v = [0, 0, 0, 0, 0, 0]$
- For $\alpha = \varepsilon = 10^{-12}$, $O_{min} = \{1\}$ therefore $x_{min} = 1$; $O_{max} = \{1, 2, 3, 4, 5\}$ therefore $x_{max} = 5$. The vector is updated to $v = [0, \varepsilon, \varepsilon, \varepsilon, \varepsilon, \varepsilon]$
- For $\alpha = .505$, $O_{min} = \{1\}$ therefore $x_{min} = 1$; $O_{max} = \{1, 2, 3, 4\}$ therefore $x_{max} = 4$. The vector is updated to $v = [0, .505, .505, .505, .505, \varepsilon]$
- For $\alpha = 1$, $O_{min} = \{1, 2\}$ therefore $x_{min} = 2$; $O_{max} = \{1, 2, 3\}$ therefore $x_{max} = 3$. The vector is updated to $v = [0, .505, 1, 1, .505, \varepsilon]$

Approximate counting always leads to an underestimation of the exact membership function of the fuzzy interval.

Acknowledgement

This research was partially supported by the Canada-Italy Innovation Award 2016 granted by the Government of Canada.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.mex.2019.10.001.

References

[1] C.C. Aggarwal, P.S. Yu, A survey of uncertain data algorithms and applications, IEEE Trans. Knowl. Data Eng. 21 (5) (2009) 609–623, doi:http://dx.doi.org/10.1109/tkde.2008.190 Institute of Electrical and Electronics Engineers (IEEE).
[2] R. Kruse, E. Schwecke, J. Heinsohn, Uncertainty and Vagueness in Knowledge Based Systems: Numerical Methods, Springer Science & Business Media, 1991.
[3] Didier Dubois, Possibility theory and statistical reasoning, Comput. Stat. Data Anal. 51 (1) (2006) 47–69, doi:http://dx.doi.org/10.1016/j.csda.2006.04.015 Elsevier BV.
[4] D. Dubois, H. Prade, The three semantics of fuzzy sets, Fuzzy Sets Syst. 90 (2) (1997) 141–150.
[5] I. Couso, D. Dubois, L. Sánchez, Random sets and random fuzzy sets as ill-perceived random variables, SpringerBriefs in Computational Intelligence, Springer, 2014.
[6] Corrado Mencar, Witold Pedrycz, Granular counting of uncertain data, Fuzzy Sets Syst. (2019), doi:http://dx.doi.org/10.1016/j.fss.2019.04.018 (available online).
[7] Lotfi A. Zadeh, The information principle, Inf. Sci. 294 (February) (2015) 540–549, doi:http://dx.doi.org/10.1016/j.ins.2014.09.026 Elsevier BV.