Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1035–1040
DOI: 10.15439/2023F3575
ISSN 2300-5963 ACSIS, Vol. 35

On Gower Similarity Coefficient and Missing Values

Marzena Kryszkiewicz
0000-0003-4736-4031
Warsaw University of Technology,
Institute of Computer Science,
Nowowiejska 15/19,
00-665 Warsaw, Poland
Email:
Marzena.Kryszkiewicz@pw.edu.pl

***Abstract*—The Gower similarity coefficient is a popular measure for comparing objects with possibly mixed-type attributes and missing values. One of its characteristics is that it calculates the coefficient value without considering attributes with missing values. In this article, we explore the properties of the coefficient in detail, including the consequences of omitting attributes with missing values. We also introduce strict lower and upper bounds on the actual similarity value on an attribute and strict lower and upper bounds on the actual value of the Gower similarity coefficient, derive a number of their properties and propose a new coefficient as a solution to the identified problem with the Gower similarity coefficient.**

***Index Terms*—Gower similarity coefficient, mixed-type attributes, quantitative attributes, qualitative attributes, dichotomous attributes, missing values.**

Introduction

THE Gower similarity coefficient [4] is a popular measure for comparing objects with possibly mixed-type attributes (quantitative, qualitative and/or dichotomous) and missing values. One of its characteristics is that it calculates the coefficient value without considering attributes with missing values. The approach is easy and intuitive and finds many applications (see, e.g. [1], [2], [3], [5], [6], [8]). It is also considered as an easily extensible template of calculating (dis)similarities of objects with mixed-type attributes [2], [5], [7]. However, as we show in the article, Gower similarity coefficient has some deficiencies. In particular, we show that in the case of objects with missing values, the coefficient may take a similarity value impossible to obtain with any replacement of missing values with values from the domains of attributes.

Our main contribution in the article includes:

- Introduction of strict lower and upper bounds on the actual similarity value on an attribute and strict lower and upper bounds on the actual value of the Gower similarity coefficient, which are obtainable after replacing missing values with respective attribute domain values.
- Showing that in the case of a pair of objects one of which has missing value for at least one quantitative attribute, the Gower similarity coefficient may take an incorrect value, which will be less than the lower bound on the actual value of the Gower similarity coefficient.
- Derivation of a number of properties of similarity value of objects on the attribute, the Gower similarity coefficient and the introduced bounds.
- Proposing new similarity coefficient G' as a correction of the Gower similarity coefficient, which eliminates the problem found for quantitative attributes with missing values.

The layout of the article is as follows: First, we recall the definitions of attribute value similarities, their weights and the Gower similarity coefficient, as well as introduce additional basic notions that are used throughout the article. Then, we show example objects for which the Gower similarity coefficient takes an incorrect value, caused by the occurrence of a missing value of a quantitative attribute for one of them. We also illustrate the consequences of the occurrence of missing values for qualitative and dichotomous attributes. Next, we introduce strict lower and upper bounds on the actual similarity value on an attribute and on the actual value of the Gower similarity coefficient, as well as derive a number of their properties. In addition, the coefficient G', being the modification of the Gower similarity coefficient, is proposed, which, unlike the original Gower similarity coefficient, always returns similarity values that do not exceed the presented lower and upper bounds.

Basic notions related to Gower Similarity Coefficient

Gower proposed a measure of objects' similarity, which can be applied in the case of qualitative attributes, quantitative attributes, dichotomous attributes or their mixtures [4]. In the measure, only the attributes for which it

is possible to determine their similarity are taken into account; the other are ignored. In particular, if for a pair of objects, an attribute value for at least one of these objects is missing, then the two objects are treated as not comparable on this attribute and the Gower similarity coefficient is calculated without taking this attribute into account.

In the remainder of the article, we assume that objects are characterized by n, where $n \geq 1$, attributes whose domains contain at least two different values. The missing value will be denoted by *. The value of attribute i of object u will be denoted by u_i.

The function $w_i(.,.)$ is used to indicate whether two objects are comparable on attribute i or not. Let u and v are objects under consideration. If u and v are comparable on attribute i, then $w_i(u,v) = 1$; otherwise $w_i(u,v) = 0$. We already mentioned that two objects u and v are incomparable on attribute i if the value of at least one of the objects is missing and so, $w_i(u,v) = 0$. However, in the case of a dichotomous attribute (indicating whether or not a feature is present), the objects may also be incomparable, even if their values are known (this happens when two objects do not have the feature represented by the dichotomous attribute).

The Gower similarity coefficient [4] for objects u and v is denoted by $G(u,v)$ and is defined as follows:

$$G(u,v) = \frac{\sum_{i=1}^{n} w_i(u,v) \times s_i(u,v)}{\sum_{i=1}^{n} w_i(u,v)},$$

where $s_i(u,v)$ is a coefficient determining similarity of two objects on attribute i, $i = 1..n$, taking values from the interval $[0,1] \cup \{undefined\}$. It is assumed that whenever $w_i(u,v) = 0$, then $w_i(u,v) \times s_i(u,v) = 0$. Thus, the Gower similarity coefficient is the average similarity of two objects on the attributes on which they are comparable.

In the case when the values of attribute i are not missing for both objects u and v, then $w_i(u,v)$ and coefficient $s_i(u,v)$ are determined as follows:

- If attribute i is qualitative:
  - $w_i(u,v) = 1$,
  - $s_i(u,v) = \begin{cases} 1, \text{if } u_i = v_i \\ 0, \text{if } u_i \neq v_i \end{cases}$;
- If attribute i is quantitative:
  - $w_i(u,v) = 1$,
  - $s_i(u,v) = 1 - \frac{|u_i - v_i|}{range_i}$;

  where $range_i = max_i - min_i$, where max_i is the maximal value of attribute i, while min_i is the minimal value of attribute i.
- If attribute i is dichotomous:
  - $w_i(u,v) = \begin{cases} 1, \text{ if } (u_i = +) \text{ and } (v_i = +) \\ 1, \text{ if } (u_i = +) \text{ and } (v_i = -) \\ 1, \text{ if } (u_i = -) \text{ and } (v_i = +) \\ 0, \text{ if } (u_i = -) \text{ and } (v_i = -) \end{cases}$,
  - $s_i(u,v) = \begin{cases} 1, \text{ if } (u_i = +) \text{ and } (v_i = +) \\ 0, \text{ if } (u_i = +) \text{ and } (v_i = -) \\ 0, \text{ if } (u_i = -) \text{ and } (v_i = +) \\ 0, \text{ if } (u_i = -) \text{ and } (v_i = -) \end{cases}$.

In the case when the value of attribute i is missing for at least one of the objects u or v, then $w_i(u,v)$ and the coefficient $s_i(u,v)$ is determined for any type of attribute i in the same way as follows:

- $w_i(u,v) = 0$,
- $s_i(u,v) = undefined$.

Now, we are ready to formally define *comparable* and *incomparable objects on an attribute*. *Objects u and v* are defined as *incomparable on attribute i* if:

- either the value of attribute i is missing for at least one the two objects
- or attribute i is dichotomous and the values of both objects are equal to $-$.

Otherwise, *objects u and v* are *comparable on attribute i*.

Property 1.

a) Objects u and v are incomparable on attribute i iff $w_i(u,v) = 0$.
b) Objects u and v are comparable on attribute i iff $w_i(u,v) = 1$.
c) If objects u and v are incomparable on attribute i, then $w_i(u,v) \times s_i(u,v) = 0$.
d) If objects u and v are comparable on attribute i, then $w_i(u,v) \times s_i(u,v) = s_i(u,v)$.
e) $s_i(u,v) = s_i(v,u)$ and $w_i(u,v) = w_i(v,u)$.

In the remainder of the article, we will use the following notation:

- *CMP_ATT*(u,v) denotes the set of attributes on which u and v are comparable; that is, *CMP_ATT*(u,v) = {attribute i| $w_i(u,v) = 1$}.
- *INCMP_ATT*(u,v) denotes the set of attributes on which u and v are not comparable; that is, *INCMP_ATT*(u,v) = {attribute i| $w_i(u,v) = 0$}.
- *INCMP*_ATT*(u,v) denotes the set of attributes on which either u or v or both have missing values.
- *INCMP^d_ATT*(u,v) denotes the set of dichotomous attributes on which both u and v have value $-$.

Property 2.

a) $G(u,v) = \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v)}{|CMP\_ATT(u,v)|}$.
b) $|CMP\_ATT(u,v)| + |INCMP\_ATT(u,v)| = n$.
c) $INCMP^*\_ATT(u,v) \cap INCMP^d\_ATT(u,v) = \emptyset$.
d) $INCMP\_ATT(u,v) = INCMP^*\_ATT(u,v) \cup INCMP^d\_ATT(u,v)$.
e) $|CMP\_ATT(u,v)| + |INCMP^*\_ATT(u,v)| \leq n$.

Objects u and v are defined as *comparable* if they are comparable on at least one attribute; that is, if $\sum_{i=1}^{n} w_i(u,v) > 0$ (or equivalently, if $|CMP\_ATT(u,v)| > 0$).

Otherwise, *objects u and v are* defined as *incomparable*; that is, when $\sum_{i=1}^{n} w_i(u,v) = 0$ (or equivalently, if $|CMP\_ATT(u,v)| = 0$). Please note that the value of $G(u,v)$ is not defined for incomparable objects. Otherwise, if u and v are comparable, then $G(u,v) \in [0, 1]$.

NEW RESULTS

A. What's Wrong with Gower Similarity Coefficient?

Though Gower similarity coefficient is appreciated by the ease and intuitiveness of dealing with attributes on which objects are incomparable, we will show that it may take an unacceptable value if the values of attributes are missing (see Example 1).

Example 1. Table I presents Set 1 of example objects characterized by qualitative attribute 1 (*colour of hair*) and quantitative attribute 2 (*age*). Let $max_2 = 100$, $min_2 = 0$, so $range_2 = 100$.

Objects u and v are comparable and different on attribute 1 (so, $w_1(u,v) = 1$ and $s_1(u,v) = 0$) and are not comparable on attribute 2 (so, $w_2(u,v) = 0$, $s_2(u,v) = undefined$). Hence, $G(u,v) = (1 \times 0 + 0 \times undefined) / (1 + 0) = 0 / 1 = 0$.

TABLE I.
SET 1 OF EXAMPLE OBJECTS

object o	1 (*colour of hair*)	2 (*age*)	$w_2(u,o)$	$s_2(u,o)$	$G(u,o)$
u	brown	**40**	1	1–|50–50|/100=1	2/2=1
v	blond	*	**0**	***undefined***	**0/1=0**
v_1	blond	0	1	1–|40–0|/100=0.6	0.6/2=0.3
v_2	blond	10	1	1–|40–10|/100=0.7	0.7/2=0.35
v_3	blond	20	1	1–|40–20|/100=0.8	0.8/2=0.4
v_4	blond	30	1	1–|40–30|/100=0.9	0.9/2=0.45
v_5	blond	**40**	**1**	**1–|40–40|/100=1**	**1/2=0.5**
v_6	blond	50	1	1–|40–50|/100=0.9	0.9/2=0.45
v_7	blond	60	1	1–|40–60|/100=0.8	0.8/2=0.4
v_8	blond	70	1	1–|40–70|/100=0.7	0.7/2=0.35
v_9	blond	80	1	1–|40–80|/100=0.6	0.6/2=0.3
v_{10}	blond	90	1	1–|40–90|/100=0.5	0.5/2=0.25
v_{11}	blond	**100**	**1**	**1–|40–100|/100=0.4**	**0.4/2=0.2**

Now we will consider what would be the Gower similarity coefficient of objects u and v_i, where v_i represents v after replacing its missing value of attribute 2 with some value from the domain range [0, 100]. Objects v_1, …, v_{11} in Table I represent object v under assumption that its actual value of attribute 2 is 0, 10, …, 100, respectively. Clearly, each instance v_i of object v is comparable with u on both attributes and is different from u on attribute 1, which is qualitative (so similarity of v_i to u on attribute 1 equals 0). Hence, $G(u,v_i) = (1 \times 0 + 1 \times s_2(u, v_i)) / (1 + 1) = s_2(u,v_i) / 2$.

Clearly, $G(u,v_i)$ reaches maximum for the greatest value of $s_2(u,v_i)$. This happens for object v_5, for which $s_2(u,v_5) = 1$ and, in consequence, $G(u,v_5) = 0.5$.

$G(u,v_i)$ reaches minimum for the least value of $s_2(u,v_i)$ (that is, for the largest absolute value of the difference between *age* of u and v_i). This happens for object v_{11}, for which $s_2(u,v_{11}) = 0.4$ and so, $G(u,v_{11}) = 0.2$. Please note that this least achievable value of 0.2 of $G(u,v_i)$ is greater than $G(u,v)$, which equals 0.

As shown in Example 1, $G(u,v)$ may take a value that is not obtainable for any actual completions of missing values of quantitative attributes of objects u and v.

In the further part of the article, we introduce strict lower and upper bounds on the actual similarity value of any objects u and v on an attribute from the set $INCMP*\_ATT(u,v)$ and on the actual value of the Gower similarity coefficient for these objects. The bounds will make it possible to check when the Gower similarity coefficient takes values unattainable for any completions of missing values.

B. Lower and Upper Bounds on Actual Similarity Value on an Attribute

Let us recall that objects u and v are not comparable on attribute i either because at least one of the objects has missing value for this attribute (i.e. $i \in INCMP*\_ATT(u,v)$) or the attribute is dichotomous and both objects have value – for it (i.e. $i \in INCMP^d\_ATT(u,v)$). If u and v are incomparable on attribute i, then $w_i(u,v) = 0$, and so attribute i does not contribute to the value of $G(u,v)$. Nevertheless, in the case of attribute $i \in INCMP*\_ATT(u,v)$, u and v may become comparable on attribute i if the actual values of attribute i become known for both objects. Then, $w_i(u,v)$ can become equal to 1, and so, $s_i(u, v)$ can contribute to the value of $G(u,v)$. Example 1 illustrates how replacing missing value of quantitative attribute i affects the values of $w_i(u, v)$, $s_i(u, v)$ and $G(u,v)$. This influence is also illustrated for a qualitative attribute and a dichotomous attribute in Examples 2 and 3, respectively.

Example 2. Table II presents Set 2 of example objects characterized by qualitative attribute 1 (*colour of hair*) and quantitative attribute 2 (*age*). Let $max_2 = 100$, $min_2 = 0$, so $range_2 = 100$.

Objects u and v are not comparable on attribute 1 ($w_1(u,v) = 0$ and $s_1(u,v) = undefined$) and are comparable on attribute 2 ($w_2(u,v) = 1$, $s_2(u,v) = 0.9$). Hence, $G(u,v) = (0 \times undefined + 1 \times 0.9) / (0 + 1) = 0.9 / 1 = 0.9$.

TABLE II.
SET 2 OF EXAMPLE OBJECTS

object o	1 (*colour of hair*)	2 (*age*)	$w_1(u,o)$	$s_1(u,o)$	$G(u,o)$
u	**brown**	40	1	1	2/2=1
v	*	30	**0**	***undefined***	**0.9/1=0.9**
v_1	**brown**	30	**1**	**1**	**1.9/2=0.95**
v_2	**blond**	30	**1**	**0**	**0.9/2=0.45**

Objects v_1 and v_2 in Table II present instances of object v after replacing its missing value of attribute 1 with some value from the domain of this attribute. Clearly, unlike v, v_1 and v_2 are comparable with u on attribute 1. Since, u and v_1 have identical value of attribute 1, their similarity on this attribute is the greatest possible; namely, $s_1(u,v_1) = 1$. Since,

u and v_2 differ on attribute 1, their similarity on this attribute is the least possible; namely, $s_1(u,v_2) = 0$. Please note that $G(u,v) \in [G(u,v_2), G(u,v_1)] = [0.45, 0.95]$.

Example 3. Table III presents Set 3 of example objects characterized by qualitative attribute 1 (*colour of hair*), quantitative attribute 2 (*age*) and dichotomous attribute 3 (*has a car*). Let $max_2 = 100$, $min_2 = 0$, so $range_2 = 100$.

Objects u and v are comparable on attributes 1 and 2 ($w_1(u,v) = w_2(u,v) = 1$, $s_1(u,v) = 0$, $s_2(u,v) = 0.9$) and are not comparable on attribute 3 ($w_3(u,v) = 0$, $s_3(u,v) = undefined$). Hence, $G(u,v) = (1 \times 0 + 1 \times 0.9 + 0 \times undefined) / (1 + 1 + 0) = 0.9 / 2 = 0.45$.

TABLE III.
SET 3 OF EXAMPLE OBJECTS

object o	1 (*colour of hair*)	2 (*age*)	3 (*has a car*)	$w_3(u,o)$	$s_3(u,o)$	$G(u,o)$
u	brown	40	**–**	0	0	2/2=1
v	blond	30	*	**0**	***undefined***	**0.9/2=0.45**
v_1	blond	30	**–**	**0**	**0**	**0.9/2=0.45**
v_2	blond	30	**+**	**1**	**0**	**0.9/3=0.3**

Objects v_1 and v_2 in Table III present instances of object v after replacing its missing value of attribute 3 with either – or +. Since, both u and v_1 have value – of attribute 3, they are not comparable on this attribute (so, $w_3(u,v_1) = 0$) and $s_3(u,v_1) = 0$. This means that attribute 3 does not contribute to the value of $G(u,v_1)$ even though its value is known both for u and v_1. Now, since, u and v_2 have values – and +, respectively, on attribute 3, they are comparable on attribute 3 (so, $w_3(u,v_1) = 1$) and their similarity on this attribute is the least possible; namely, $s_3(u,v_1) = 0$. Please note that $G(u,v) \in [G(u,v_2), G(u,v_1)] = [0.3, 0.45]$.

In Examples 1, 2 and 3, we considered instances of example object u, with known values for all attributes, and object v, with missing value only for one given attribute i. We considered all or some instances of object v in which missing value was replaced by possible actual values including those instances of object v whose similarity on attribute i was the least and greatest, respectively. Clearly, these least and greatest values are lower and upper bounds, respectively, on similarity values of objects u and v on the examined attributes.

Let $i \in INCMP^*\_ATT(u, v)$. *Lower bound on the actual similarity value of u and v on attribute i* will be denoted by $\underline{s}_i(u,v)$, while *upper bound on the actual similarity value of u and v on attribute i* will be denoted by $\overline{s}_i(u,v)$. The associated *weights for the bounds* will be denoted as $\underline{w}_i(u,v)$ and $\overline{w}_i(u,v)$, respectively.

In Table IV, we provide the values of the similarity bounds $\underline{s}_i(u,v)$ and $\overline{s}_i(u,v)$ and their weights, respectively, under assumption that the value of attribute i is missing for at least one object. In fact, $\underline{s}_i(u,v) = \underline{s}_i(v,u)$, $\underline{w}_i(u,v) = \underline{w}_i(v,u)$, $\overline{s}_i(u,v) = \overline{s}_i(v,u)$ and $\overline{w}_i(u,v) = \overline{w}_i(v,u)$, thus, without loss of generality, we assume that the value of attribute i is missing for object v. The results are provided for quantitative, qualitative and dichotomous attributes. We also indicate for which possible actual values of v and eventually u, $s_i(u,v) = \underline{s}_i(u,v)$ and $s_i(u,v) = \overline{s}_i(u,v)$, respectively. Thus, we show that $\underline{s}_i(u,v)$ and $\overline{s}_i(u,v)$ are strict lower and upper bounds on the actual similarity value of objects u and v on each attribute $i \in INCMP^*\_ATT(u, v)$.

Please note that $\underline{w}_i(u,v)$ equals 1 for each attribute $i \in INCMP^*\_ATT(u, v)$. On the other hand, the lower bound $\underline{s}_i(u,v) = 0$ in all cases considered in Table IV except for quantitative attribute i whose value is missing for only one of the two compared objects. In that exceptional case, $\underline{s}_i(u,v)$ depends on the known value of the other object and can be

TABLE IV.
STRICT SIMILARITY BOUNDS $\underline{s}_i(u,v)$, $\overline{s}_i(u,v)$ AND THEIR ASSOCIATED WEIGHTS $\underline{w}_i(u,v)$ AND $\overline{w}_i(u,v)$ FOR MISSING VALUE OF OBJECT v AND KNOWN OR MISSING VALUE OF OBJECT u.

Type of attribute *i*	**Value of attribute *i* for object *u***	$\underline{w}_i(u,v)$	$\underline{s}_i(u,v)$	**When $s_i(u,v) = \underline{s}_i(u,v)$?**	$\overline{w}_i(u,v)$	$\overline{s}_i(u,v)$	**When $s_i(u,v) = \overline{s}_i(u,v)$?**
qualitative	missing	1	0	when actual value of v is different from actual value of u	1	1	when actual value of v is equal to actual value of u
	x	1	0	when actual value of v is different from x	1	1	when actual value of v is equal to x
quantitative	missing	1	0	when actual value of v is minimal and actual value of u is maximal or vice versa	1	1	when actual value of v is equal to actual value of u
	x	1	$\min\{(x - min_i), (max_i - x)\} / range_i$	when the absolute value of the difference between x and actual value of v is the largest possible; that is, is equal to $max\{(x - min_i), (max_i - x)\}$.	1	1	when actual value of v is equal to x
dichotomous	missing	1	0	when actual value of v is different from actual value of v	1	1	when actual value of v and actual value of u are equal to +
	+	1	0	when actual value of v is equal to –	1	1	when actual value of v is equal to +
	–	1	0	when actual value of v is equal to +	0	0	when actual value of v is equal to –

greater than 0 (as shown in Table IV, in this case, $\underline{s}_i(u,v) = \min\{(x - min_i), (max_i - x)\} / range_i$).

Property 3. Let i be a quantitative attribute. Let the value of attribute i be missing for object v and be equal to x for object u. Then:

a) $\underline{s}_i(u,v)$ reaches maximum, which is equal to 0.5, for $x = (min_i + min_i) / 2$.
b) $\underline{s}_i(u,v)$ reaches minimum, which is equal to 0, for $x = min_i$ or $x = max_i$.

Proof: Follows from $\underline{s}_i(u,v)$ for a quantitative attribute (see Table IV).

Note also that for each attribute $i \in INCMP*\_ATT(u, v)$, upper bound $\overline{s}_i(u,v) = 1$ and $\overline{w}_i(u,v) = 1$, unless attribute i is dichotomous and its value is equal to $-$ for one object, say u, and is missing for the other object, say, v. In that exceptional case, $\overline{w}_i(u,v) = 0$ and $\overline{s}_i(u,v) = 0$ (which corresponds to the situation when the actual value of v is also equal to $-$), while $\underline{w}_i(u,v) = 1$ and $\underline{s}_i(u,v) = 0$ (which corresponds to the situation when the actual value of v equals +). In the former case, attribute i does not contribute to the Gower similarity coefficient, while in the latter case, attribute i contributes to it with the least possible value of 0.

C. Lower and Upper Bounds on Actual Value of Gower Similarity Coefficient

We start with defining lower and upper bounds on the actual value of the Gower similarity coefficient, which are achievable after replacing all missing values in the compared objects with some values from the domains of corresponding attributes.

Lower bound on the actual value of $G(u,v)$ is denoted by $\underline{G}(u,v)$ and is defined as follows:

$$\underline{G}(u,v) = \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v) + \sum_{i \in INCMP*\_ATT(u,v)} \underline{w}_i(u,v) \times \underline{s}_i(u,v)}{|CMP\_ATT(u,v)| + \sum_{i \in INCMP*\_ATT(u,v)} \underline{w}_i(u,v)}.$$

Upper bound on the actual value of $G(u,v)$ is denoted by $\overline{G}(u,v)$ and is defined as follows:

$$\overline{G}(u,v) = \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v) + \sum_{i \in INCMP*\_ATT(u,v)} \overline{w}_i(u,v) \times \overline{s}_i(u,v)}{|CMP\_ATT(u,v)| + \sum_{i \in INCMP*\_ATT(u,v)} \overline{w}_i(u,v)}.$$

Clearly, if $|CMP\_ATT(u,v)| + \sum_{i \in INCMP*\_ATT(u,v)} \underline{w}_i(u,v) > 0$, then $\underline{G}(u,v)$ is the strict lower bound on the actual value of $G(u,v)$, which is obtainable for some completion of missing attribute values of objects u and v, while $\overline{G}(u,v)$ is the strict upper bound on the actual value of $G(u,v)$ provided $|CMP\_ATT(u,v)| + \sum_{i \in INCMP*\_ATT(u,v)} \overline{w}_i(u,v) > 0$.

As shown in Table IV, the weight $\underline{w}_i(u,v) = 1$ for each attribute $i \in INCMP*\_ATT(u,v)$. Hence, $\underline{G}(u,v)$ can be rewritten as presented in Property 4:

Property 4.

a) $\underline{G}(u,v) = \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v) + \sum_{i \in INCMP*\_ATT(u,v)} \underline{s}_i(u,v)}{|CMP\_ATT(u,v)| + |INCMP*\_ATT(u,v)|}$.

b) $\underline{G}(u,v) \geq \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v) + \sum_{i \in INCMP*\_ATT(u,v)} \underline{s}_i(u,v)}{n}$ if $|CMP\_ATT(u,v)| + |INCMP*\_ATT(u,v)| > 0$.

Proof: Ad a) By definition of $\underline{G}(u,v)$ and the fact that $\underline{w}_i(u,v) = 1$ for each attribute $i \in INCMP*\_ATT(u,v)$ (see Table IV).
Ad b) By Property 4a and Property 2e.

Property 5. If there are no quantitative attributes in $INCMP*\_ATT(u, v)$, then

a) $\underline{G}(u,v) = \frac{\sum_{i \in CMP_{ATT(u,v)}} s_i(u,v)}{|CMP_{ATT(u,v)}| + |INCMP*\_ATT(u,v)|}$.

b) $\underline{G}(u,v) \geq \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v)}{n}$ if $|CMP\_ATT(u,v)| > 0$.

c) $\underline{G}(u,v) \leq G(u,v)$ if $|CMP\_ATT(u,v)| > 0$.

Proof: Ad a) By Property 4a and the fact that $\underline{s}_i(u,v) = 0$ for each non-quantitative attribute i in $INCMP*\_ATT(u, v)$ (see Table IV).
Ad b) By Property 5a and Property 2e.
Ad c) By Property 5a and Property 2a.

Example 4. Let us consider again objects u and v from Example 1 (see also Table I), whose attribute 2 is quantitative. Then, $G(u,v) = 0$, $\underline{s}_2(u,v) = \min\{(40 - 0), (100 - 40)\} / 100 = 0.4$ (see Table IV), $\underline{G}(u,v) = (1 \times 0 + 1 \times 0.4) / (1 + 1) = 0.2$. Thus, $\underline{s}_2(u,v) > G(u,v)$ and $\underline{G}(u,v) > G(u,v)$.

Example 4 allows us to conclude what follows:

Property 6. Let u and v be comparable objects. Let i be a quantitative attribute with missing value for object u and known value for object v. Then:

a) It is probable that $\underline{s}_i(u,v) > G(u,v)$.
b) If $\underline{s}_i(u,v) > G(u,v)$, then it is probable that $\underline{G}(u,v) > G(u,v)$.

Corollary 1. It is probable that $\underline{G}(u,v) > G(u,v)$ when there is a missing value in u or v. If $\underline{G}(u,v) > G(u,v)$, then $G(u,v)$ takes an incorrect value, which cannot be obtained for any possible actual value of attribute i of object u.

To avoid the problem stated in Corollary 1, one may use, depending on an application, the lower bound $\underline{G}(u,v)$, the upper bound $\overline{G}(u,v)$ or an appropriately modified version of $G(u,v)$ instead of $G(u,v)$ itself. Below we introduce new $G'(u,v)$ similarity coefficient defined as follows:

$$G'(u,v) = \frac{\sum_{i \in CMP\_ATT(u,v)} s_i(u,v) + \sum_{i \in INCMP*\_ATT(u,v), \underline{s}_i(u,v) > G(u,v)} \underline{s}_i(u,v)}{|CMP\_ATT(u,v)| + |\{i \in INCMP*\_ATT(u,v) | \underline{s}_i(u,v) > G(u,v)\}|}.$$

In fact, $G'(u,v)$ can be regarded as an improved version of $G(u,v)$.

Let $INCMP*\_QNT\_ATT(u,v)$ be the set of the quantitative attributes in $INCMP*\_ATT(u,v)$. Now, we will express $G'(u,v)$ in terms of attributes in $CMP\_ATT(u,v) \cup INCMP*\_QNT\_ATT(u,v)$.

Property 7. Let $|CMP\_ATT(u,v)| > 0$. Then:

$$G'(u,v) = \frac{\sum_{i\in CMP\_ATT(u,v)} s_i(u,v)+\sum_{i\in INCMP*\_QNT\_ATT(u,v,} \underline{s}_i(u,v)}{|CMP\_ATT(u,v)|+ \ |\{i\in INCMP*\_QNT\_ATT(u,v)|\ \underline{s}_i(u,v)>G(u,v)|}.$$

Proof: By assumption, u and v are comparable, so $G(u,v) \geq 0$. If i is a qualitative or dichotomous attribute in $INCMP*\_ATT(u, v)$, then $\underline{s}_i(u,v) = 0$ (see Table IV), and so, $\underline{s}_i(u,v)$ is not greater than $G(u,v)$. So, $\underline{s}_i(u,v)$ can be greater than $G(u,v)$ only if i is a quantitative attribute in $INCMP*\_ATT(u, v)$; i.e., if $i \in INCMP*\_QNT\_ATT(u,v)$.

Please note that $G'(u,v)$ differs from $G(u,v)$ in that the value of $G'(u,v)$ is calculated not only on the attributes on which u and v are comparable (as in the case of $G(u,v)$), but also on those quantitative attributes i on which u and v are not comparable provided $\underline{s}_i(u,v) > G(u,v)$.

Property 8. Let u and v be comparable objects. Then:

a) $G'(u,v) \geq G(u,v)$.
b) $G'(u,v) \geq \underline{G}(u,v)$.

Proof: Ad a) By definition of $G'(u,v)$ and Property 2a.
Ad b) By definition of $G'(u,v)$ and Property 4a.

Example 5. In the case of objects u and v from Example 1 (see also Table I), $G'(u,v) = \underline{G}(u,v) = 0.2 > G(u,v) = 0$.

We will consider now the properties of the upper bound on Gower similarity coefficient.

Property 9. If there are no dichotomous attributes in $INCMP*\_ATT(u, v)$, then:

$$\overline{G}(u,v) = \frac{\sum_{i\in CMP\_ATT(u,v)} s_i(u,v)+|INCMP*\_ATT(u,v)|}{|CMP\_ATT(u,v)|+ |INCMP*\_ATT(u,v)|}.$$

Proof: By definition of $\overline{G}(u,v)$ and the fact that $\overline{s}_i(u,v) = 1$ and $\overline{w}_i(v,u) = 1$ for any non-dichotomous attribute i on which u and v are incomparable (see Table IV).

Finally, we check the relationship between $\overline{G}(u,v)$ and $G'(u,v)$ as well as between $\overline{G}(u,v)$ and $G(u,v)$.

Property 10. Let u and v be comparable objects. Then:

a) $\overline{G}(u,v) \geq G'(u,v)$.
b) $\overline{G}(u,v) \geq G(u,v)$.

Proof: Ad a) In the proof, we will use the property saying that $\overline{s}_i(u,v) = 1 \geq \underline{s}_i(u,v)$ and $\overline{w}_i(v,u) = 1$ for any attribute $i \in INCMP*\_QNT\_ATT(u,v)$ (*) and that for any attribute $j \in INCMP*\_ATT(u,v) \setminus INCMP*\_QNT\_ATT(u,v)$ either: (i) $\overline{s}_j(u,v) = 1$ and $\overline{w}_j(u,v) = 1$ or (ii) $\overline{w}_j(u,v) = 0$ (**).

Thus, by definition,

$$G'(u,v) =$$

$$\frac{\sum_{i\in CMP\_ATT(u,v)} s_i(u,v)+\sum_{i\in INCMP*\_QNT\_ATT(u,v,} \underline{s}_i(u,v)}{|CMP\_ATT(u,v)|+ \ |\{i\in INCMP*\_QNT\_ATT(u,v)|\ \underline{s}_i(u,v)>G(u,v)|}$$

/ by (*) /

$$\leq \frac{\sum_{i\in CMP\_ATT(u,v)} s_i(u,v)+\sum_{i\in INCMP*\_QNT\_ATT(u,v)} \overline{w}_i(u,v)\times\overline{s}_i(u,v)}{|CMP\_ATT(u,v)|+ \sum_{i\in INCMP*\_QNT\_ATT(u,v)} \overline{w}_i(u,v)}$$

/ by (**) /

$$\leq \frac{\sum_{i\in CMP\_ATT(u,v)} s_i(u,v)+\sum_{i\in INCMP*\_ATT(u,v)} \overline{w}_i(u,v)\times\overline{s}_i(u,v)}{|CMP\_ATT(u,v)|+ \sum_{i\in INCMP*\_ATT(u,v)} \overline{w}_i(u,v)}$$

$$= \overline{G}(u,v).$$

Ad b) By Property 10a and Property 8a.

SUMMARY

In the article, we introduced lower and upper bounds on the actual similarity value on an attribute and on the actual value of the Gower similarity coefficient. We showed that the Gower similarity coefficient for two objects may take an incorrect value, which would be less than the lower bound on the actual value of the Gower similarity coefficient for those objects, if one of the objects has a missing value for at least one quantitative attribute. To solve this problem, we introduced coefficient G', being a modification of the Gower similarity coefficient, that is free from this deficiency. A number of properties of similarity value of objects on the attribute, the Gower similarity coefficient, the introduced lower and upper bounds and the coefficient G' were derived.

REFERENCES

[1] B. Ben Ali, Y. Massmoudi, "K-means clustering based on gower similarity coefficient: A comparative study," 2013 5th International Conference on Modeling, Simulation and Applied Optimization, ICMSAO 2013. https://doi.org/10.1109/ICMSAO.2013.6552669.

[2] S. S. K J. Chae and W. Y. Yang, "Cluster analysis with balancing weight on mixed-type data," The Korean Communications in Statistics, vol. 13, no. 3, 2006, pp. 719–732, http:\\DOI:10.5351/CKSS.2006.13.3.719.

[3] J. Fontecha, R. Hervás, and J. Bravo, "Mobile Services Infrastructure for Frailty Diagnosis Support based on Gower's Similarity Coefficient and Treemaps," Mobile Information Systems, vol. 10, Article ID 728315, 20 pages, 2014. https://doi.org/10.1155/2014/728315.

[4] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties, Biometrics, " Vol. 27, No. 4. (Dec., 1971), pp. 857-871, https://doi.org/10.2307/2528823.

[5] S. Pavoine, J. Vallet, A.-B. Dufour, S. Gachet, and H. Daniel, "On the challenge of treating various types of variables: application for improving the measurement of functional diversity," Oikos, 118(3) 2009, pp. 391-402, https://doi.org/10.1111/j.1600-0706.2008.16668.x.

[6] G. Philip and B. S. Ottaway, "Mixed data cluster analysis: an illustration using cypriot hooked-tang weapons, " Archaeometry, vol. 25, no. 2, 1983, pp. 119–133, https://doi.org/10.1111/j.1475-4754.1983.tb00671.x.

[7] J. Podani and D. Schmera: "Generalizing resemblance coefficients to accommodate incomplete data," Ecological Informatics 66 (2021) 101473, https://doi.org/10.1016/j.ecoinf.2021.101473

[8] G. Tuerhong and S. B. Kim, "Gower distance-based multivariate control charts for a mixture of continuous and categorical variables," Expert Systems with Applications, 41(4 PART 2), 2014, pp. 1701–1707, https://doi.org/10.1016/j.eswa.2013.08.068.