Set-theoretic similarity measures

Maria Rifqi, Bernadette Bouchon-Meunier

Set-theoretic similarity measures

M. Rifqi and B. Bouchon-Meunier
LIP6 – Pôle IA
8, rue du Capitaine Scott – 75015 Paris, France
{Maria.Rifqi, Bernadette.Bouchon-Meunier}@lip6.fr

Abstract. In this paper, we give an overview of general classes of similarity measures which presents the advantage of being inserted in a cognitive framework and being available for objects described by means of non-classical kinds of values, such as linguistic values, as well as traditional numerical attributes. We also give means to prioritize one measure over the others.

1 Introduction

Similarity is a widely used concept, with utilizations in various fields, such as pattern recognition, case-based reasoning, image processing, approximate reasoning, machine learning, information retrieval for instance. There exist many definitions of similarity or resemblance, or conversely, dissimilarity or distance. In the case where the objects to be compared are not defined on a metrical universe, distances are clearly not appropriate. Several types of similarity are nevertheless available [1], which have been connected with seminal works in psychology [6].

2 Knowledge representation

Let us consider descriptions of objects defined on a universe Ω. The elements of Ω can be, for instance, sets of symbols (words in text mining, features in images...), intervals of the set of real numbers $\mathbb{R}$ (measurements,...), fuzzy sets of a reference set U (representation of linguistic descriptions in database querying,...). In this last case, the membership function of A is denoted by f_A. We suppose that an *order* $\subseteq$ is defined on Ω, and an operation of *difference* $-$ on Ω is such that: $A - B$ is monotonous with respect to A, according to $\subseteq$, and $A \subseteq B$ implies $A - B = \emptyset$.

We also define operations of *union* $\cup$ and *intersection* $\cap$ on Ω. We finally suppose given a mapping $M : \Omega \rightarrow R^+$ such that: $M(\emptyset) = 0$ and M is monotonous with respect to $\subseteq$.

For instance, in the case of a set Ω of symbols or intervals of $\mathbb{R}$, the operations are classical intersection, union and difference, and $M(A)$ is the cardinality $|A|$ of A for symbols, its norm $\|A\|$ for intervals. The order $\subseteq$ is the classical inclusion.

Let us finally consider a reference set U, the set Ω of fuzzy sets of U, the order $\subseteq$ is the classical inclusion of fuzzy sets ($A \subseteq B$ if and only if $f_A \leq f_B$), the intersection and union

are defined by $\min$ and $\max$ operators. *Differences* $-$ of fuzzy sets can be:

$$f_{A-B}(x) = \max(0, f_A(x) - f_B(x)) \text{ [7]} \quad (1)$$

$$f_{A-_2B}(x) = \begin{cases} f_A(x) & \text{if } f_B(x) = 0 \\ 0 & \text{if } f_B(x) > 0 \end{cases} \quad (2)$$

The mapping M is a *fuzzy set measure*, for instance: $M_1(A) = \int_U f_A(x)dx$, $M_2(A) = \sup_{x \in U} f_A(x)$, $M_3(A) = \sum_{count} f_A(x)$ if U is countable.

3 Measures of similitude

Let us now consider measures of comparison of elements of Ω [1].

An *M-measure of similitude* (`m.sim`) on Ω is a mapping : $S : \Omega \times \Omega \rightarrow [0, 1]$, defined as: $S(A, B) = F_S(M(A \cap B), M(B - A), M(A - B))$, for a given mapping $F_S : R^{+3} \rightarrow [0, 1]$, such that $F_S(u, v, w)$ is non-decreasing in u, and non-increasing in v and w.

Obviously, the notion of M-measure of similitude is still very general and corresponds to mappings with very different behaviors. In order to help choosing one of them in a particular problem solving, we consider the following additional properties which may be satisfied by M-measures of similitude.

- *reflexivity* ($S(A, A) = 1$ for any A) which means that $F_S(u, 0, 0) = 1$ for any $u \neq 0$.
- *exclusiveness* ($S(A, B) = 0$ for any A and B such that $A \cap B = \emptyset$), which means that $F_S(0, v, w) = 0$ for any v and w different from 0.
- *symmetry* ($S(A, B) = S(B, A)$ for any A and B), which means that $F_S(u, v, w) = F_S(u, w, v)$ for any u, v, w.

We then distinguish the following `m.sim` of particular interest :

- An *M-measure of satisfiability* (`m.sat`) is an exclusive and reflexive `m.sim` independent of the third component: $S(A, B) = F_S(M(A \cap B), M(B - A))$, for a function $F_S : R^{+2} \rightarrow [0, 1]$ such that $F_S(u, v) = F_S(u, v, .)$. As a consequence, an `m.sat` satisfies the *containment* property (if $B \subseteq A$, $B \neq \emptyset$, then $S(A, B) = 1$).
- An *M-measure of resemblance* (`m. res`) is a symmetric and reflexive `m.sim`: $S(A, B) = F_S(M(A \cap B), M(B - A), M(A - B))$.

4 Discrimination power of measures of similitude

4.1 Measures of satisfiability

A `m.sat` corresponds to a situation in which we consider a reference object or a class and we need to decide if a new object is compatible with it or satisfies it. For instance, `m.sat` are appropriate for rule base systems.

4.1.1 Scale sensitivity

It is desirable that a `m.sat` depends only on the relative weights of its components and not on the scale of the system. In order to obtain an objective measure, we propose [4] to normalize the `m.sat`. We note: $X = M(A \cap B)$ and $Y = M(B - A)$. We consider: $x = \frac{X}{\sqrt{X^2+Y^2}}$, the reduced intersection and $y = \frac{Y}{\sqrt{X^2+Y^2}}$, the reduced distinctive feature.

As $x^2 + y^2 = 1$, the domain of definition of the `m.sat` is a quarter of circle. It can be described by a unique argument ϕ, with $\phi = \arctan \frac{y}{x}$. We note the `m.sat` $S(A, B) = \eta(\phi)$, with η decreasing with respect to ϕ, such that $\eta(\frac{\pi}{2}) = 0$ and $\eta(0) = 1$.

We can represent (figure 1) the reference set A by the vector $V_{\vec{A}}$ and the set B by a vector $V_{\vec{B}}$. When the two vectors are orthogonal, then the satisfiability vanishes: $S(A, B) = 0$. More generally, the satisfiability appears as a projection, and a lack of satisfiability is represented as a deviation in figure 1.

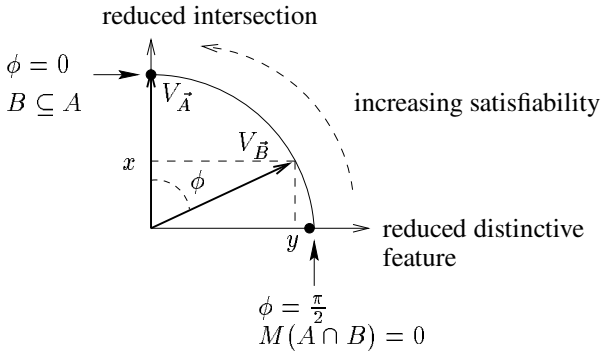

Figure 1: New representation of a measure of satisfiability

This new form of a `m.sat`, expressed by a unique variable, has the advantage of not being dependent upon the size of the system. Furthermore, this normalization makes the definition of a `m.sat` more simple insofar as the argument is a segment $[0, \frac{\pi}{2}]$ and not a quarter of plan.

4.1.2 Examples

There are of course many possible choices for the satisfiability measure η. Among them, let us distinguish the following forms (see figure (a) of table 1):

- $\eta_1(\phi) = 1 - \frac{2}{\pi}\phi$. It is a linear satisfiability function.
- $\eta_2(\phi) = \cos\phi$, where $\eta_2(\phi)$ can be seen as the scalar product $V_{\vec{A}} \cdot V_{\vec{B}}$.
- $\eta_3(\phi) = \frac{1}{1+\tan\phi}$.
- $\eta_4(\phi) = 1 - \sin\phi$.

The third measure can be rewritten as: $S(A, B) = \frac{M(A \cap B)}{M(B)}$, in the case of fuzzy sets of U, with $M = M_3$. The last one can be rewritten as: $S(A, B) = 1 - M(B - A)$ introduced in [2] for fuzzy sets with $M = M_2$ and with the difference $-_2$.

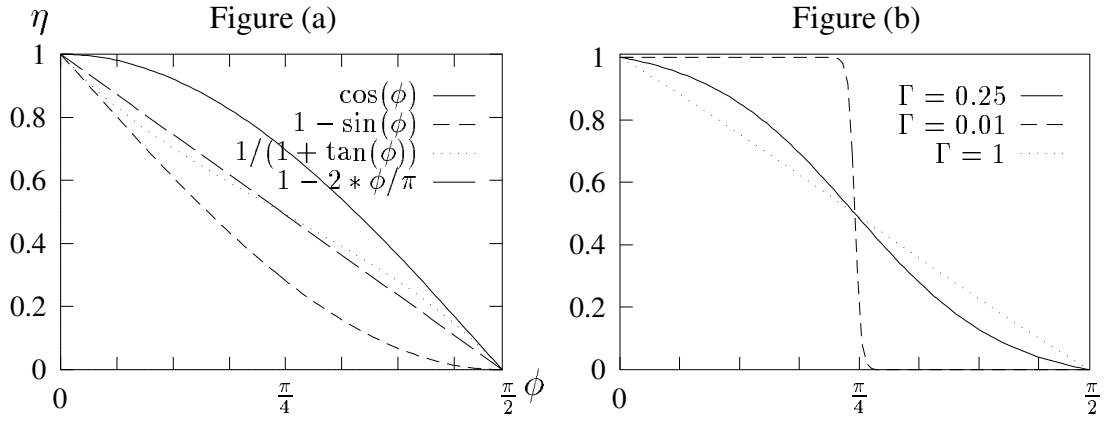

Table 1: The behaviour of measures of satisfiability

4.1.3 Discrimination power

With this new representation, the `m.sat` can be easily compared. We consider the *discrimination power* of a measure of satisfiability as given by the derivative $\eta'(\phi)$ of η.

For every possible η, we have: $\int_0^{\frac{\pi}{2}} \eta'(\phi)d\phi = -1$. This means that the total discrimination power $\eta'(\phi)$ has to be distributed on the $[0, \frac{\pi}{2}]$ interval, but a high discrimination power somewhere implies a low discrimination power elsewhere. Accordingly, it is necessary to choose a measure with a discrimination power suitable for the considered application. This suggests a method of construction of a `m.sat`.

We can remark that no function with a high discrimination power for $\eta(\phi) = 1/2$ but a low discrimination for $\eta(\phi) = 0$ and $\eta(\phi) = 1$ is available in figure (a) of table 1. Nevertheless, this kind of measures means that if a description is not far from the reference, then the satisfiability is near 1 because the difference is not significative. If a description is very far from the reference, we can consider that the satisfiability is null. We propose such an interesting measure based on the Fermi-Dirac function $F_{FD}(\phi) = \frac{1}{1+\exp^{\frac{(\phi-\frac{\pi}{4})}{\Gamma}}}$, where $\Gamma \in I\!R^+$ controls the decrease of the curve. This measure is defined as:

$$\eta(\phi) = \frac{F_{FD}(\phi) - F_{FD}(\frac{\pi}{2})}{F_{FD}(0) - F_{FD}(\frac{\pi}{2})}$$

The choice of Γ enables to define a `m.sat` more or less severe, as shown on the figure(b) of table 1.

4.2 Measures of resemblance

A `m.res` is used for a comparison between the descriptions of two objects, of the same level of generality, to decide if they have many common characteristics.

Measures of resemblance are appropriate for case-based reasoning or instance-based learning. In clustering methods, distances can be replaced by a `m.res`. More generally, similarity-based classification methods [5] have to use `m.res` as soon as all objects have the same level of generality. Let us denote $Z = M(A - B)$.

We focus on `m.res` satisfying the property of *exclusiveness*.

Following our normalization procedure, we define: $x = \frac{X}{\sqrt{X^2+Y^2+Z^2}}$, $y = \frac{Y}{\sqrt{X^2+Y^2+Z^2}}$, $z = \frac{Z}{\sqrt{X^2+Y^2+Z^2}}$, for $(X, Y, Z) \neq (0, 0, 0)$. Similarly to the case of `m.sat`, this ensures that an exclusive `m.res` is not dependent on the scale of the problem.

The domain of study is now restricted to a part of the unity sphere since $x^2 + y^2 + z^2 = 1$. Geometrically, the sphere is simply obtained by a rotation of the satisfiability circle around the x-axis (see figure 2). The vector representation is still valid.

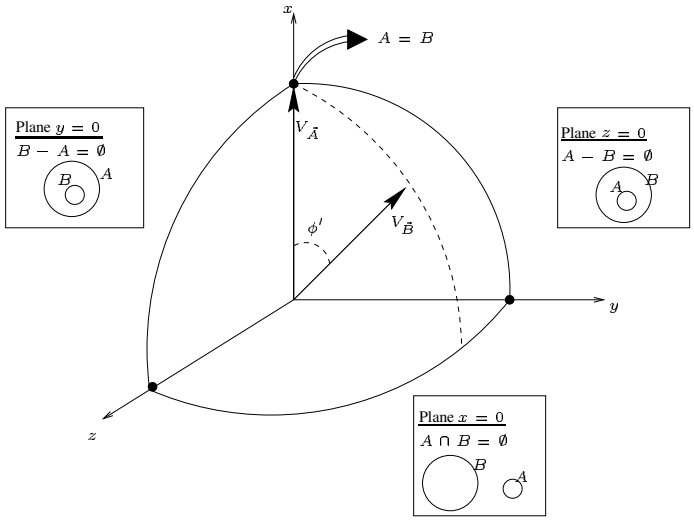

Figure 2: New representation of an exclusive measure of resemblance

Let us consider $\rho = \xi(y, z)$ with ξ non decreasing with regard to x and z, $\xi(0, 0) = 0$ and $\xi(y, z) = \xi(z, y)$. This means that ρ can be described by any symmetrical function with respect to y and z. Let us define $\psi = \arctan(\frac{\rho}{x})$ and a `m.res` $S(A, B) = \nu(x, \rho)$ such that: $\nu(x, \rho)$ is increasing with respect to x and decreasing with respect to ρ, $\nu(0, \rho) = 0$ if $\rho \neq 0$, $\nu(x, 0) = 1 \ \ \forall x$.

These conditions show that the problem has been reduced to a satisfiability measure. We can therefore use again the solution described in the preceding section dealing with satisfiability. With this definition of ρ, an exclusive resemblance appears as a satisfiability where a global distinctive feature ρ is defined by $\rho = \xi(y, z)$, from the two individual distinctive features y and z.

We can also consider different exclusive `m.res` as we have already done with `m.sat`. In the case where $\rho^0 = y + z$, we get $\nu_0 = \frac{1}{1+\frac{\rho^0}{x}} = \frac{1}{1+\tan\psi}$. This measure corresponds to the `m.sat` η_3. Furthermore, ν_1 can be also written as: $S(A, B) = \frac{M(A\cap B)}{M(A\cup B)}$ with M such that: $M(A \cup B) = M(A \cap B) + M(A - B) + M(B - A)$, for instance M_3. This measure was introduced in [3].

Other definitions of ρ can be envisaged, for instance: $\rho' = \sqrt{y^2 + z^2}$ or $\rho'' = (\sqrt{y} + \sqrt{z})^2$ associated with ν' and ν''. The choice of a particular form of ρ has an effect on the measure of resemblance because this parameter represents distinctive elements. We can notice that, $\rho'' \geq \rho^0 \geq \rho' \ \ \forall\ y, z$. As ρ has a decreasing effect on an exclusive `m.res`, the above relation implies that, for a given x and for all y and z, $\nu''(x, \rho'') \leq \nu(x, \rho^0) \leq \nu'(x, \rho')$. This relation means that $\nu''(x, \rho'')$ penalizes more the differences between two sets than $\nu(x, \rho^0)$ and that $\nu(x, \rho^0)$ penalizes more the differences than $\nu'(x, \rho')$. Furthermore, a particular ρ is sensitive

to the symmetry between y and z. Indeed, if distinctive features y and z are unbalanced for instance, it means that $y >> z$ or $z >> y$, the behaviours of two given ρ can be opposite and then the order of resemblances of objects are totally inverted.

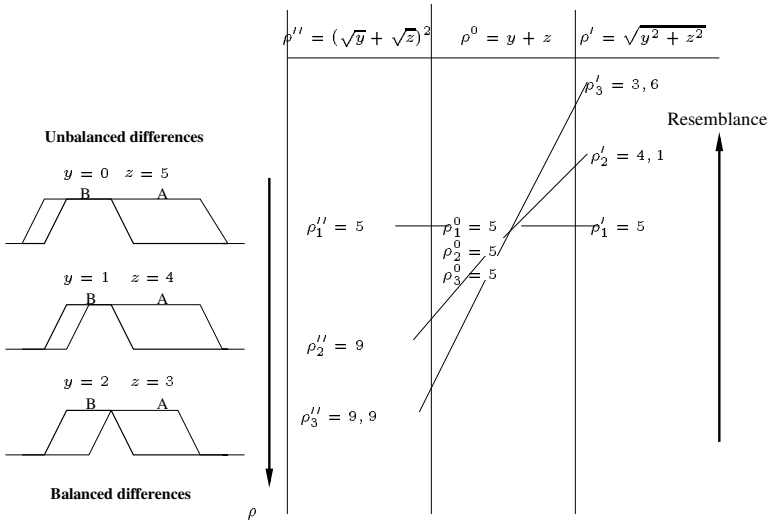

Figure 3: The effects of different definitions of ρ on exclusive measures of resemblance.

5 Conclusion

In the domain of non metric similarity measures, the paper focuses on those based on sets (fuzzy or classical). Even in this restricted study, the choice of a similarity remains very large. We have proposed a way to focalize on particular family of measures in two steps: the first one consists in distinguishing general properties such as symmetry, reflexivity, exclusiveness,etc. Then, the second step enables to refine the family found in the first step by describing the discrimination power of the desired measure.

References

[1] B. Bouchon-Meunier, M. Rifqi, and S. Bothorel. Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84(2):143–153, 1996.

[2] B. Bouchon-Meunier and L. Valverde. Analogy relations and inference. In *Proceedings of* 2^{nd} *IEEE International Conference on Fuzzy Systems*, pages 1140–1144, San Fransisco, 1993.

[3] D. Dubois and H. Prade. *Fuzzy Sets and Systems, Theory and Applications*. Academic Press, New- York, 1980.

[4] M. Rifqi, V. Berger, and B. Bouchon-Meunier. Discrimination power of measures of comparison. *Fuzzy Sets and Systems*, 110(2):189–196, March 2000.

[5] M. Rifqi, S. Bothorel, B. Bouchon-Meunier, and S. Muller. Similarity and prototype based approach for classification of microcalcifications. In 7^{th} *IFSA World Congress*, pages 123–128, Prague, 1997.

[6] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[7] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, 8 and 9:199–249, 301–357, 43–80, 1975.