ICCBR 2007, 13-16 August 2007, Belfast, Northern Ireland

SIMILARITY, ANALOGY AND CASE-BASED REASONING

Bernadette Bouchon-Meunier
LIP6, Université Pierre et Marie Curie-Paris 6, 104 avenue du Président Kennedy
75016 Paris, France
bernadette.bouchon-meunier@lip6.fr

1. Introduction

Similarity has been pointed out as a key concept is a number of domains, such as linguistics, semiology, psychology and computational intelligence. Similarities are very useful for all attempts to construct "human-like" automated systems, since human beings are very efficient in using them to deal with real world complexity. For instance, Rissland [17] stresses on their importance in artificial intelligence, arguing that concepts are often "messy", with a core of easy to classify instances and a "penumbra" of other instances which can be managed on the basis of similarities with others. Similar elements are regarded from different points of view by using resemblances, distances, closeness, proximity, synonymy, dissimilarities, analogies…

Fuzzy sets are closely related to the definition of similarities because of their capacity to represent subjective information, resulting from real world complexity, and gray areas of interpretation as presented by Rissland and because of the graduality inherent in their definition, in agreement with the natural behavior of intuitive similarities. We propose an overview of similarities in the framework of fuzzy logic, similarity measures enabling the user to preserve the flexibility and graduality human beings have in mind when they deal with similarities, and use expressions such as "very similar", "rather similar" "more similar than"…

We would like to stress on the fact that many of the measures, methods and properties which are pointed out can be used in a general environment, not necessarily involving a fuzzy set based representation, the fuzziness and graduality appearing in the only similarities themselves. There exist various types of similarity measures: for binary data, for fuzzy data, for numerical data, for structured data… We will focus on classical definitions related to binary data and their extensions to fuzzy data in a general formalization incorporating most of the well known measures, classifying them and proposing new ones, to help the user to choose one of them according to the problem he is coping with. For similarities regarding precise numerical data, see [4].

This presentation is owing to Maria Rifqi, Marcin Detyniecki and Marie-Jeanne Lesot's work.

2. Similarities in a fuzzy framework

We first consider classical definitions of similarity measures, such as Jaccard, Dice or Ochiai coefficients. For an overview of such definitions, see for instance [12]. Tversky's seminal work on similarities in psychology [19] has provided a formal definition of measures of similarities as a linear combination of the measures of their common and distinctive features. In an attempt to approach the flexibility of the concept of similarity for human beings, and extending Zadeh's original definition of similarity [20] in a fuzzy framework, we have proposed a general definition of so-called measures of comparison involving fuzzy descriptions of features [1], encompassing classical coefficients as particular cases and being in agreement with Tversky's so-called contrast model.

More precisely, let Ω be a given universe and $F(\Omega)$ the set of its fuzzy sets. On the basis of a fuzzy set measure $M : F(\Omega) \to \Re^+$, and two operations on $F(\Omega)$, namely a difference Θ and an intersection $\cap$, we define a measure of comparison on Ω as a mapping $S : F(\Omega) \times F(\Omega) \to [0,1]$ such that $S(A, B) = F_S\big(M(A \cap B), M(B\Theta A), M(A\Theta B)\big)$, for a mapping $F_S : \Re^3 \to [0,1]$.

The choice of a measure of comparison by a user depends on his needs with regard to the problem he has to solve. Types of measures of comparison have been identified to help him to choose an appropriate measure. Particular cases are:

- measures of *resemblance*, which are reflexive and symmetrical, increasing in $M(A \cap B)$, decreasing in $M(A\Theta B)$ and $M(B\Theta A)$.
- measures of *satisfiability*, which are reflexive, exclusive, and independent of $M(A\Theta B)$, not necessarily symmetrical, increasing in $M(A \cap B)$, decreasing in $M(B\Theta A)$.
- measures of *inclusion*, reflexive, exclusive and independent of $M(B\Theta A)$, not necessarily symmetrical, increasing in $M(A \cap B)$, decreasing in $M(A\Theta B)$.

All of them evaluate how similar the elements are. Measures of resemblance represent the "similarity" between elements of the same kind or level and can be used for instance in clustering or data mining, while satisfiability and inclusion measures evaluate the "similarity" of a new element with a reference. A satisfiability measure evaluates to which extent B is compatible with A and it can be used in decision trees or case-based reasoning, for instance. An inclusion measure evaluates to which extent B can be considered as a particular case of A and it is useful when working on databases or semantic networks, for instance.

Other particular cases are measures of *dissimilarity*, independent of $M(A\Theta B)$, non decreasing in $M(B\Theta A)$ and $M(A\Theta B)$, indicating the degree of difference between objects.

For more details about this framework, see [12].

3. Properties of measures of comparison

Further properties of these measures have been pointed out to refine the classification of comparison measures and provide more help to the user.

Their *discrimination power* has been for instance studied in [14][16] to evaluate the influence of a variation of data on the value of the comparison measure and observe if a small difference between values of the considered variable causes a severe variation of the measure of comparison or not. The so-called Fermi-Dirach measure has been introduced and studied on that occasion because of its interesting properties.

Equivalence between comparison measures is another interesting question for users who need a ranking of elements according to their resemblance with a given case, rather than the precise value of the similarity for each element. An equivalence class of comparison measures contains measures providing the same ranking, whatever the values of the considered variables are. The so-called "equivalence in order" of two comparison measures can be proven to be identical with the "equivalence by a function", indicating that one comparison measure can be expressed from the other one by means of a strictly increasing function. This property is easy to check by means of graphical representations of the measures. "Equivalence in order" can also be proven to be identical with the "equivalence in level sets", showing that, if we define a *threshold* used to select elements similar to a given case at least at this level, we can find the threshold providing the same selection with another comparison measure of the same equivalence class. See [11] for more details.

As an example, we can mention that Jaccard, Dice and Fermi-Dirach measures are equivalent and provide the same ranking, whatever the values of variables are. Ochiai measure is not in this class and provides different results when looking for "similar" elements to a given case.

Aggregation of comparison measures is also interesting for all cases described by means of several variables or attributes. It can be proven [3] that using triangular norms or OWA operators for the aggregation of values of measures of resemblance, satisfiability, inclusion or dissimilarity for all attributes preserves their properties and provides global measures of the same kind.

4. Utilizations of measures of comparison for case-based reasoning.

We can identify various utilizations of measures of comparison for case-based reasoning. They can obviously be used to identify elements similar to a studied case. They can help to obtain a clustering of cases which have been retrieved or all cases of the database, to improve the efficiency of the retrieval. They can also be used for the construction of prototypes as abstract representatives of classes of situations or concepts, in agreement with Rosch’s work on typicality in psychology [18]. Exceptions can also be taken into account and recognized as informative by special forms of similarity-based clustering. More details about their utilization in clustering can be found in [4][5][6], in the construction of prototypes in [8][9]. Examples of applications of similarities are given in [6][9][10][13][15];

5. Conclusion

Similarities are clearly key concepts for automated processes understanding human beings subtlety and acquiring their perceptive capability of categorization. They appear to be very useful for several steps of case-based reasoning and they provide a richer “toolbox” to the user than classical distances or cosine coefficient.

References

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

2. B. Bouchon-Meunier, C. Marsala, M. Rifqi, Interpolative reasoning based on graduality, *Proceedings 9th IEEE Int. Conf. on Fuzzy Systems*, San Antonio, USA, pp. 483-487 (2000).
3. B. Bouchon-Meunier et M. Rifqi, OWA operators and an extension of the contrast model, in The *Ordered Weighted Averaging Operators: Theory, Methodology, and Applications*, R. R. Yager and J. Kacprzyk (Eds.), Kluwer Academic Publishers, pp. 29-35, 1997.
4. M.-J. Lesot. Similarity, typicality and fuzzy prototypes for numerical data. In *Res-Systemica*, 5 (Special issue on the 6th European Congress on Systems Science, Paris 2005), 2005.
5. M.-J. Lesot. Typicality-based clustering. *Int. Journal of Information Technology and Intelligent Computing*, 1(2) pp. 279-292, 2006.
6. M.-J. Lesot et R. Kruse. Data summarisation by typicality-based clustering for vectorial data and nonvectorial data. *IEEE Conference on Fuzzy Systems* (Fuzz-IEEE'06) Vancouver, Canada, 2006.
7. M.-J. Lesot et R. Kruse. Gustafson-Kessel-like clustering algorithm based on typicality degrees. *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'06)*, Paris, France, 2006
8. M.-J. Lesot L. Mouillet et B. Bouchon-Meunier, Fuzzy prototypes based on typicality degrees. *Proc. of Fuzzy Days 04*, Springer, Advances on Soft Computing, pp. 125-138, Dortmund, Allemagne, 2006
9. M.-J. Lesot, M. Rifqi, B. Bouchon-Meunier, Fuzzy prototypes: from a cognitive view to a machine learning principle, *in* H. Bustince, F. Herrera, J. Montero (eds.) *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models Studies*, Springer Verlag, pp. 431-453, 2007.
10. J.-F. Omhover, M. Detyniecki, and B. Bouchon-Meunier, A Region-Similarity-Based Image retrieval system, *in* B. Bouchon-Meunier, G. Coletti, R. Yager (Eds.), *Modern Information Processing: From Theory to Applications,* Elsevier, 2005.
11. J.-F. Omhover, M. Detyniecki, M. Rifqi, and B. Bouchon-Meunier, Image Retrieval using Fuzzy Similarity: measure equivalence based on invariance in ranking, *Proceedings of the IEEE International Conference on Fuzzy Systems - FUZZ-IEEE'2004*, pp. 1367-1372, Budapest, Hungary, 2004.
12. M. Rifqi. Mesures de comparaison, typicalité et classication d'objets flous : théorie et pratique. *PhD thesis, Université Paris VI*, 1996.
13. M. Rifqi, Constructing prototypes from large databases, *Proc. International Conference IPMU'96*, Granada, pp. 301-306, 1996.
14. M. Rifqi, V. Berger, B. Bouchon-Meunier, Discrimination power of measures of comparison, *Fuzzy Sets and Systems* 110, 2, pp. 189-196, 2000.
15. M. Rifqi, S. Bothorel, B. Bouchon-Meunier, S. Muller, Similarity and prototype based approach for classification of microcalcifications, *International Fuzzy Systems Association (IFSA'97)*, Prague, 1997.
16. M. Rifqi, M. Detyniecki and B. Bouchon-Meunier, Discrimination power of measures of resemblance, *International Fuzzy Systems Association (IFSA'03)*, Istanbul, 2003.
17. E. Rissland, AI and similarity, *IEEE Intelligent Systems*, Vol.21, pp. 39-49, 2006.
18. E. Rosch, Principles of categorization, *in* E. Rosch and B. Lloyd (eds.) *Cognition and categorization*, 27-48, Lawrence Erlbaum,1978.
19. A. Tversky, Features of similarity, *Psycho. Rev.* 84, 4, pp. 327-352,1977.

20. L. A. Zadeh, Similarity relations and fuzzy ordering, *Information Science*, pp.177-200, 1971.