Similarities in Fuzzy Data Mining: From a Cognitive View to Real-World Applications

Bernadette Bouchon-Meunier, Maria Rifqi, and Marie-Jeanne Lesot

UPMC Univ Paris 06, CNRS,UMR 7606, LIP6, F-75005, Paris, France
{bernadette.bouchon-meunier, maria.rifqi,
marie-jeanne.lesot}@lip6.fr

Abstract. Similarity is a key concept for all attempts to construct human-like automated systems or assistants to human task solving since they are very natural in the human process of categorization, underlying many natural capabilities such as language understanding, pattern recognition or decision-making. In this paper, we study the use of similarities in data mining, basing our discourse on cognitive approaches of similarity stemming for instance from Tversky's and Rosch's seminal works, among others. We point out a general framework for measures of comparison compatible with these cognitive foundations, and we show that measures of similarity can be involved in all steps of the data mining process. We then focus on fuzzy logic that provides interesting tools for data mining mainly because of its ability to represent imperfect information, which is of crucial importance when databases are complex, large, and contain heterogeneous, imprecise, vague, uncertain or incomplete data. We eventually illustrate our discourse by examples of similarities used in real-world data mining problems.

Keywords: similarity, data mining, categorization, prototype, fuzzy sets.

1 Introduction

Since similarities are very natural in the human process of categorization underlying many natural capabilities such as language understanding, pattern recognition or decision-making, they are naturally fundamental for all attempts to construct human-like automated systems or assistants to human task solving, and particularly in data mining and information retrieval.

Those domains are difficult to cope with for various reasons. First, most of the databases are complex, large, and contain heterogeneous, imprecise, vague, uncertain or incomplete data. Furthermore, the queries may be imprecise or subjective in the case of information retrieval; the mining results must be easily understandable by a user.

Fuzzy logic provides interesting tools for such tasks, mainly because of its capability to manage imprecise categories, to represent imperfect information, for instance by means of fuzzy sets, graduality, measures of resemblance or aggregation methods.

J.M. Zurada et al. (Eds.): WCCI 2008 Plenary/Invited Lectures, LNCS 5050, pp. 349–367, 2008.

We propose to explore the capabilities of similarities to interact with fuzzy methods in several steps of the data mining process, information retrieval or knowledge discovery, such as clustering, construction of prototypes, utilization of expert or association rules or fuzzy querying, for instance.

With this object, we present a view of the concept of similarity rooted in cognitive psychology, and we discuss its utilization in the framework of fuzzy data mining. The paper is organized as follows: in Section 2, we consider the cognitive point of view on similarity and the related categorization notion, pointing out the elements that could be of interest in data mining. In Section 3, we examine the use of similarity in the data mining framework, underlining its central role in all steps of this process. In Section 4, we focus on the case of fuzzy logic: after recalling existing measures, we describe a general framework for comparison measures that is compatible with the cognitive foundations, and state properties that can be desired from similarity measures. These properties provide guides for selecting a measure appropriate for a given problem. In Section 5, we eventually present some real-world applications where these paradigms have been exploited among others to manage various types of data, such as image retrieval or risk analysis.

2 Similarity and Categorization in Cognitive Science

Similarities, in a general sense, have been widely studied in cognitive psychology, from different points of view, and in particular for the categorization task. The latter aims at reducing the amount of information in order to decrease our cognitive effort and at reflecting the structure of the real world [1][2]. Data mining issues are connected to these objectives: likewise, it aims at extracting, from large data bases, relevant information that still reflects the structure of the whole base.

If bridges have been made between data mining and cognitive science regarding similarity, many aspects studied by one of the communities remain unknown by the other one. In this section, we would like to point out some of the elements raised on the subject of similarities that could be of interest in data mining and related topics. Our purpose in this brief cursory glance at similarities in cognitive psychology is to show the various possible angles we can choose to treat them and to leave doors open to new visions of similarities in the data mining domain.

The concepts of categorization, similarities and prototypes are intrinsically connected, even though their relationships are not uniformly accepted and various approaches of these concepts intertwine. For the sake of simplicity, we consider them successively, pointing some of their interrelations.

2.1 Categorization

In psychology, several approaches of categorization [3] can be distinguished, according to the underlying structure of the categories they assume: one approach assumes that there exist rules used for the recognition of categories; another one is based on the knowledge of properties shared by members of a category. A third one considers that categories are based on the recognition of similarities.

One vision of categorization supposes an all-or-none relationship between categories and objects: an object belongs to a category or it does not. This way, categories are defined in terms of necessary conditions, and not of similarities. For instance [4] points out the existence of categories that are based on explicit definitions, such as the category of triangles defined by a list of geometric properties, or on an ad hoc process gathering objects for a given purpose, such as Valentine's day gifts.

Under the assumption that similarities are the basis of natural categorization, there exists a variety of points of view. The notion of family resemblance introduced by Wittgenstein [5] in 1953 (for instance the family of games) does not use an explicit definition of the similarity involved in the categorization but a subjective judgment, which may be dependent on the context.

Exemplar models [4] can be regarded along the same lines and consider that categories are represented in terms of individual instances. A new object or element is then classified in a category if it is more similar to elements already stored in this category than to elements of other categories. This theory does not take into account any representative of a category.

On the opposite, another stream dealing with family resemblance considers a category by means of a central representative of the category called a prototype, and a graded structure around it [2], formed by objects similar to the prototype. Furthermore, a category can be associated to several prototypes in case of a diversity of subcategories.

More precisely, Rosch [2] considers that an object can better represent its category than another one. The typicality of an object for a given category depends on its resemblance to the other members of the category and its differences to the members of other categories. In other words they are spread on a scale, or a gradient, of typicality: the more typical an object, the more attributes it shares with other members of its category and the less attributes it shares with members of the other categories.

Kleiber [6] extends Rosch's approach, underlining the notion of fuzzy frontiers and the fact that the belonging to a category is based on the degree of similarity with the prototype of this category. Furthermore, the notions of typicality degree and membership degree are clearly distinct whereas they were not in Rosch's approach. It means that, even if an object is less typical of a category than another one, it does not necessarily belong to this category to a smaller extent: although an ostrich is not a typical bird, it still totally belongs to the bird category. This example moreover highlights the fact a category can be binary defined and still characterized in terms of typicality.

In the studies regarding similarity and typicality, there have been several attempts to prove that a differentiation can be made between them [3]. One reason is that frequency of instantiation can be regarded as involved in the identification of typicality in addition to similarity. The extreme option considers that frequency is the most important factor in typicality [7]. Familiarity with exemplars may appear to be involved in the construction of a graded structure of a category [8], and familiarity is both related to frequency and context. Another reason to differentiate similarity and typicality deals with causality, and takes into account the variability and the existence of changes in the identification of categories, for instance related to the history of changes, with a consideration of time in the identification of categories [9].

It can be considered that human beings naturally form concepts through this prototype mechanism. For instance Posner [10] considers that prototypes play a part in the formation of abstract ideas, taking into account the variability of instances expressed in terms of distances between patterns.

A different view of categories can be based on intrinsic coherence, regarded as the existence of links between properties that constitute a kind of conceptual core [4]. This view is not incompatible with the notion of typicality but it describes atypical elements by means of the non-existence or co-existence of some properties. This leads to the identification of hybrid categories and does not accept any graduality. There is no explicit reference to similarities in such a theory. Category variability [11] is pointed out as a motivation for this different vision of categorization, since properties satisfied by a category may depend on the context.

2.2 Similarity

As it is accepted that most natural categories are structured in terms of family resemblance or centered around prototypes, similarity plays a central role in category structure.

The seminal work by Tversky [12] rejects the classic assumption of the need of a metric space to define similarities. He assumes in particular that symmetry is not a necessity for a similarity judgment, since an observed object can be compared to a reference object in an asymmetric way due to the status of the two objects. He also rejects the necessity of the transitivity property. He introduces the so-called contrast model, defining a measure of similarity of two objects as a function increasing with respect to the features common to these objects and decreasing with respect to their distinctive features. He suggests more properties to require from measures of similarity, and he mentions the importance of context, reducing it to a choice of features. He observes that similarity has two faces: the first one is causal in that it serves as a basis for the classification of objects, and the second one is derivative as it is influenced by the existing classification.

He proposes the so-called ratio model as a particular case of the contrast model, defining similarity measures by the following form, for two given non-negative parameters α and β :

$$s_{\alpha,\beta}(A, A') = \frac{f(A \cap A')}{f(A \cap A') + \alpha f(A \Theta A') + \beta f(A' \Theta A)} \quad (1)$$

It is to be noted that Tversky considers features as basic granules of the description of objects: for instance, for the description of a human face, a feature can be the presence or absence of a smiling mouth, a frowning mouth or a straight eyebrow. A particular case corresponds to features considered as attributes with values in universes of discourse, for instance "mouth" with values {smiling, frowning, neutral}.

A more shaded approach [13] suggests that one can observe a difference between similarity judgments and categorization tasks when deeper features than perceptual elements are used for the categorization and this difference could be rubbed out if several levels of similarity were taken into account, from perceptual similarities to

conceptual similarities based on domain theories. This approach does not seem to have been much explored.

Recently, attempts to take into account changes and variations of categories in time have given rise to a dynamic similarity processing formalization [14] as opposed to the classical static similarities we described above. Several views of dynamic similarities are possible [15], either concerned by the history of perceptual patterns, or approaching previsions of categories expected in the future, for instance through an adaptation process.

2.3 Related Concepts

There exist various interpretations of the general concept of similarity; words like similarity, analogy, proximity or closeness are often used in an undifferentiated way, even though they refer to different definitions.

In a cognitive sense, analogy is formalized in a simple representation by "as A is to B, so C is to D", and is based on an identity of relations between situations or objects which can be of a completely different nature, involving the idea of structure or function. On the contrary, similarity, simply expressed as "A is similar to B", identifies a resemblance between two objects. Similarity and analogy are still connected in several aspects. The most obvious connection lies on the fact that the recognition of analogy is often based on similarities. Furthermore, the so-called alignment model of similarity [16] is based on the assumption that mental representations are structured and evaluating the similarity between elements or objects takes into account relations in the structure, for instance relations between perceptual units or classic semantic relations such as meronymy or holonymy relations. This model stems from representations of analogy in addition to semantic descriptions.

The concepts of proximity and closeness are related to distance measures. In many cases, a similarity measure can be defined as the dual of a dissimilarity measure or a distance, on the basis of a rule of the form "the less distant, the more similar". Nevertheless, similarity and dissimilarity are concepts which can be considered as antinomic or complementary, depending on the angle: dissimilarities, called differences, are recognized as different from the opposite of similarities by Tversky [12], whereas in the case of prototypes, similarity and dissimilarity are two complementary components of a global approach of classification.

Inclusion is another related concept that has been mainly attached to the idea of implication or inference in cognitive psychology. It is involved in the identification of similarities, for instance in a property-based categorization [4].

3 Similarities in Data Mining

Similarities (or dissimilarities) have been widely used in artificial intelligence. Rissland [17] points out their importance and underlines their central role, explaining it by the difficulty of representing real-world concepts. She considers that real-world concepts are “messy” in the sense that they have grey areas of interpretation, which

leads to an open-textured property, they change and are submitted to a non-stationary property and they have exceptions. She suggests that representing messy concepts in artificial intelligence presents a challenge that can be braved thanks to the notion of similarity. We complete this assumption in claiming that these properties of real-world concepts lead to fuzzy–set based representation.

3.1 Standard Data Mining

The well-known description of the data mining process given by Fayyad [18] presents a succession of four steps:

(i) from databases or data warehouses, a selection process extracts relevant data,

(ii) these relevant data are submitted to cleaning or transformation operations in order to construct a training set,

(iii) on this training set, a machine learning method is used to elicit a model of information,

(iv) this model is submitted to an interpretation in order to obtain knowledge understandable from the user or expert.

Now all these four steps can benefit from the use of some types of similarities.

In step (i), the selection can be achieved thanks to a matching between query and data, on the basis of similarities.

In step (ii), data cleaning and data reduction strategies are various and similarities, among others, bring solutions to these processes. There exist for instance various approaches to the management of missing data [19], and some of them exploit the notion of similarity or distance, especially those based on the use of clusters of similar observations to assign a value replacing a missing one [20] [21]. Methods to simplify data by means of attribute selection and dimensionality reduction can also be based on the use of distances.

In step (iii), many machine learning methods can be related to the concept of similarity. Clustering is for instance based on the principle of grouping elements as close as possible to each other with regard to attribute values and also (for some methods) as far as possible from elements of other groups. Statistical techniques such as Support Vector Machines lie on kernel functions that are nothing but similarities.

Similarities are explicitly used in non-classical reasoning approaches, such as case-based reasoning, analogical reasoning, similarity-based reasoning, where they constitute the core of the methods themselves.

Inductive learning is an exemplar-based construction of rules describing categories and the similarity of instances belonging to a category is action-oriented, an action being either the identification of a class or a decision to make. In the case of decision tree construction, it can be considered that the conditional entropy used to elicit the rules corresponds to a probabilistic version of similarity.

In step (iv), the passage from abstract models to knowledge needs an interpretation phase. In this part, again, similarities can be used for several purposes, for instance for rule base simplification [22].

3.2 Similarities in Fuzzy Data Mining

If we focus on data mining in a fuzzy set theory setting, specific needs of similarity management arise.

In database management, similarities are useful to compare an approximate value involved in a query to all possible solutions stemming from the database.

In the construction of a model, fuzzy association rules can be managed thanks to similarities [23]. In fuzzy inductive learning, the discretization phase splitting the attribute values in two or more fuzzy classes is based on similarities underlying this process of categorization [24]. The choice of the best attribute in the construction of a fuzzy decision tree relies on the optimization of a measure of the discriminating power of an attribute with regard to a class: the measure of classification ambiguity [25] proposed by Yan and Shaw is for instance based on fuzzy similarities.

When fuzzy decision trees are used to classify an example, similarity is used to compare its attribute values to those associated with edges of the tree, whereas a simple binary matching step is applied if a standard decision tree is used [26].

At the final level of the interpretability, expressivity of rules can be improved by means of linguistic modifiers closely related to similarities [27][28] or for a linguistic expression of categories [29]. In the case of a model taking the form of if-then rules for instance, similarities can be used to merge several rules and to simplify the model [30].

We have presented examples of situations where measures of similarities are useful. This is the reason why we present how they are represented in a fuzzy setting.

4 Similarities in a Fuzzy Setting

From the rapid presentations of similarity and categorization issues in psychology described in Section 2, the links with fuzzy sets appear clearly. The recurrent occurrence of variability in categories and their graded structure incline us to take advantage of the graduality and flexibility inherent in fuzzy sets to define measures of similarity and to model categories.

This has obviously been achieved from the early beginning of fuzzy set theory since L.A. Zadeh [31] has introduced the concept of similarity relation as an extension of equivalence relation, that presents the advantage of providing a crisp partitioning of data on the basis of a fuzzy knowledge of their relations. It should be remarked that the introduced softness is limited, properties inherited from classic relations such as transitivity and symmetry being preserved. Attempts to go further in the flexibility have led to indistinguishability relations [32] accepting a version of transitivity less constrained than similarity relations.

Such a fuzzy relation is defined on a given universe and provides the degree of similarity of any pair of elements in this universe. For instance, if a discrete universe of colors is considered, orange is more similar to red than to blue. In the case where one wants to compare fuzzy sets of the universe rather than precise elements, we can use extensions to sets of fuzzy sets of these similarity or indistinguishability relations. For instance, compared colors are regarded as fuzzy sets of a continuous universe. In this framework, measures of similarity or resemblance have been proposed.

4.1 Measures of Similarity

Measures of similarity (or dissimilarity) have obviously been introduced out of the scope of fuzzy set theory in a wider framework. The measures proposed by Jaccard [33], Dice [34] or Ochiai [35] have been extensively used in many domains, and they belong to a set-based view of similarities, taking into account the numbers of elements common to the compared objects or distinct between them.

Many set-theoretical measures actually correspond to particular cases of Tversky ratio model [12], as defined in Equation (1), Jaccard coefficient being for instance associated with parameters $\alpha = \beta = 1$, and Dice with parameters $\alpha = \beta = 1/2$.

Besides, many distances have also been introduced in a geometric vision of the descriptions of objects to compare (e.g. Minkowski, Chebychev, Hausdorff, Mahalonobis) and they have been used in various areas.

Such measures have given rise to a variety of measures of comparison of fuzzy sets, evaluating either their resemblance or their difference. Many attempts have been made to compare them [36] and to get them into some kind of order, in such a way that they could efficiently be used in practical domains, such as image processing, pattern recognition or data mining [37][38][39][40]. Most of the proposed classifications are based on a distinction between set-theoretic and geometric procedures, and some of them add the third class of logic-based procedure [41].

A thorough analysis of the existing so-called measures of similarity points out very different forms of measures of comparison of fuzzy sets, going farther than the simple notion of similarity. The most common are distances, measures of dissimilarity and inclusion. Generic terms such as compatibility measures [41], comparison indices [42], are proposed to take into account all measures of "matching" between fuzzy sets.

Tversky's contrast model is sometimes mentioned to study set-theoretic similarity measures in a fuzzy framework [41], and fuzzy versions of Tversky's contrast model have been proposed [43][44][45] in specific areas. In the following section, we describe a different connection between Tversky's approach and similarity of fuzzy sets, in which a general classification framework for comparison measures is introduced.

4.2 General Framework for Measures of Comparison

Working with various real-world applications requiring diverse measures to evaluate how fuzzy descriptions of objects differ or are similar, we have had a double concern: first, to help to classify such measures, embracing as many kinds as possible in a unique framework, in order to propose to the user the most convenient solutions to his problem; secondly, to remain compatible with cognitive psychology views of measures of similarity. Tversky's model has been chosen because of its degree of generality and the wide spectrum of potential instantiation. So-called general measures of comparison [46][47] have been introduced, encompassing the main measures of similarity, dissimilarity, satisfiability, resemblance, inclusion.

We briefly recall the principles of this framework before showing how it has been used in practical applications.

Let Ω be a given universe and $F(\Omega)$ the set of its fuzzy sets, equipped with the classical inclusion $\subseteq$, a fuzzy set measure $M : F(\Omega) \rightarrow \Re^+$, such that such that

$M(\varnothing) = 0$ and M is monotonous with regard to $\subseteq$, and two operations on $F(\Omega)$, namely an intersection $\cap$ and a difference Θ such that $A \subseteq B$ implies $A\Theta B = \varnothing$. We define a measure of comparison on Ω as a mapping:

$$S : F(\Omega) \times F(\Omega) \rightarrow [0,1] . \tag{2}$$

of the form:

$$S(A, B) = F_S\left(M(A \cap B), M(B\Theta A), M(A\Theta B)\right), \tag{3}$$

for a mapping $F_S : \Re^3 \rightarrow [0,1]$.

It must be remarked that they only follow the basic Tversky's requirement of matching. Interesting properties may be required from measures of comparison in order to capture all necessary behaviors involved in the comparison of elements in practical applications. Reflexivity and symmetry are simple extensions of classical notions. Exclusiveness is satisfied if $S(A, B) = 0$ for any A and B such that $A \cap B = \varnothing$.

Four main types of measures of comparison are identified to help the user in various kinds of processes. Regarding similarity assessments, three processes stem from the applications, in agreement with psychological studies: either an object is compared to a reference and measures of satisfiability or inclusion are introduced, or two objects with the same status are compared and measures of resemblance are presented. We also distinguish dissimilarity from the negation of similarity, introducing so-called measures of dissimilarity. The following classes are thus defined:

- Measures of *resemblance* are reflexive and symmetrical, increasing in $M(A \cap B)$, decreasing in $M(A\Theta B)$ and $M(B\Theta A)$.
- Measures of *satisfiability* 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* are reflexive, exclusive and independent of $M(B\Theta A)$, not necessarily symmetrical, increasing in $M(A \cap B)$, decreasing in $M(A\Theta B)$.
- Measures of *dissimilarity* are independent of $M(A \cap B)$, non-decreasing in $M(B\Theta A)$ and $M(A\Theta B)$, such that $S(A, B) = 0$ for any A and B such that $A\Theta B = \varnothing$ and $B\Theta A = \varnothing$. They indicate the degree of difference between features.

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, semantic networks or relations between properties for instance. A dissimilarity measure is important for the construction of prototypes or in some clustering methods.

Measures of resemblance, satisfiability and inclusion are proved to be in agreement with Tversky's requirements of monotonicity, independence, solvability, invariance [46] for measures of similarity. Although there exist measures of similarity in Tversky's contrast model which are not in any of the above categories, the latter correspond to most of the needs in practical applications. For more details about this framework, see [46] [47][48].

Looking at Tversky's similarity measures $r_{\alpha,\beta}$ as defined in Equation (1), it can be noted that they have properties of measures of resemblance when $\alpha = \beta$, measures of satisfiability when $\alpha = 0$, and measures of inclusion when $\beta = 0$.

To point out the interest of differentiating those different measures, let us consider three particular measures, denoting f_A the membership function of a fuzzy set A:

- Measure of resemblance

$$S(A, B) = \frac{\int_\Omega f_{A\cap B}(u)du}{\int_\Omega f_{A\cup B}(u)du} \tag{4}$$

corresponding to Jaccard coefficient in a classical framework.

- Measure of satisfiability

$$S(A, B) = \frac{\int_\Omega f_{A\cap B}(u)du}{\int_\Omega f_{B}(u)du} \tag{5}$$

- Measure of inclusion

$$S(A, B) = \frac{\int_\Omega f_{A\cap B}(u)du}{\int_\Omega f_{A}(u)du} \tag{6}$$

The difference between them is very subtle for the user, and the two last ones can seem equivalent at a first glance. It must be noted that they have a very different nature and using one or the other is not neutral.

4.3 Properties of Measures of Comparison

Among all properties of measures of comparison that could be presented to help the user in his choice of one of them, we stress on two main studies providing guidelines to make a choice among the jungle of measures, concerning the discrimination power of measures on the one hand, ranking properties on the other hand.

The sensitivity of measures of comparison with respect to small variations of the compared elements is an important component in the choice of one of them. Choosing a representation of similarity measures avoiding the scaling problem [49][50], it is possible to show differences of behavior between measures very discriminating for small feature values variations, or for large feature values. Based on this study of the discriminating power of similarity measures [49], a new measure of similarity has been defined, the so-called Fermi-Dirac measure defined as follows:

$$S(A, B) = \frac{F_{FD}(\varphi) - F_{FD}(\Pi / 2)}{F_{FD}(0) - F_{FD}(\Pi / 2)} \tag{7}$$

with

$$F_{FD}(\varphi) = \frac{1}{1 + \exp\left(\frac{\varphi - \varphi_0}{\Gamma}\right)}, \quad \varphi = \arctan\left(\frac{M(B\Theta A) + M(A\Theta B)}{M(A \cap B)}\right) \tag{8}$$

$\Gamma \in \Re^+$ and $\varphi_0 \in [0, \Pi / 2]$ are parameters balancing the selectivity of Fermi-Dirac measure. It presents the particular property of being discriminating around the specific value φ_0 chosen by the user, with an intensity defined by the Γ parameter.

The second property of measures of comparison we put forward is more specific of problems where values of a similarity measure are not important as such, and where only the induced order matter [43]. This for instance occurs in the case of document retrieval systems, where the user is interested in the list of documents more similar to his request, ignoring the similarity score of each document [51]. Likewise, in case-based reasoning, the *n* first candidates are of importance, irrespective on their similarity values.

Classes of measures of resemblance providing the same ranking have been pointed out [52]. Three definitions of the equivalence of resemblance measures have been proposed, which appear to be themselves equivalent.

Given a reference object A, two resemblance measures S and S' are order equivalent if and only if any element B provides a value $S(A, B)$ greater than the value $S(A, C')$ corresponding to another object C' whenever $S'(A, B)$ is greater than $S'(A, C')$. Formally, this can be written as:

$$\forall A, B, C, \qquad S(A, B) \leq S(A, C) \Leftrightarrow S'(A, B) \leq S'(A, C). \tag{9}$$

Such a condition is equivalent to the existence of a strictly increasing function from the image of S to the image of S':

$$f : \mathrm{Im}(S) \rightarrow \mathrm{Im}(S'). \tag{10}$$

such that:

$$S' = f \circ S \tag{11}$$

Another equivalent definition of the equivalence of measures of resemblance can be considered, based on a common structure in level sets for S and S', in such a way that, for any α in the image of S, there exists a unique β in the image of S' such that the α-level set of S is identical with the β-level set of S'.

It is to be remarked that, if we use a threshold α to select all objects B best resembling A at a level at least equal to α, the collection of objects we obtain is different if we use S or S'. If S and S' are equivalent, we obtain the same collection

of objects if we consider a threshold α when using S and a threshold $\beta = f(\alpha)$ when using S'. If we fix the cardinality of the collection of objects we look for, we obtain the same collection when using equivalent resemblance measures.

If resemblance measures S and S' are not equivalent, for a given value of S', there may exist several values of S, which means that for one object resembling A at the level β for S', it is possible to find several objects resembling A when using S.

Considering the basic measures we have recalled, Jaccard (see Equation (4)) and Fermi-Dirac (see Equation (7)) measures are for instance equivalent, while Jaccard and Ochiai are not equivalent. Moreover, Tversky's measures $s_{\alpha,\beta}$ and $s_{\alpha',\beta'}$ (1) are equivalent whenever $\alpha.\beta' = \alpha'.\beta$ [51].

4.4 Similarity-Based Prototypes

Beyond similarity, the fuzzy setting makes it possible to implement other notions related to similarity and categorization in the cognitive framework. In particular, prototypes, viewed as significant representatives of the categories, can be built in agreement with the cognitive principles mentioned in Section 2.1, on the basis of similarity judgments. Since the graded character of prototypes has been emphasized, it is indeed natural to choose a fuzzy knowledge-based representation of prototypes, avoiding the choice of crisp representatives of a category such as the mean value for an attribute.

Several approaches have been proposed since the seminal one introduced by Zadeh [51], for instance based on fuzzy expected values [54] or fuzzy summaries [55]. Fuzzy prototype construction has often been considered as identical with fuzzy clustering, which is yet somewhat different. Fuzzy clustering forms categories of objects similar with regard to attribute values, while fuzzy prototypes propose abstract representatives of categories, generally not belonging to the categories of objects, associated with the most representative fuzzy value of each attribute. An extensive study of fuzzy prototypes has been presented in [56]. The principles of their construction are summarized [46][57], coherent with Rosch's vision of prototypes [1].

The basic concept is the degree of typicality of an object with respect to a category. It can be regarded as the aggregation of a degree of internal resemblance measure of an attribute value with regard to all other values of the same attribute for objects of the same category on the one hand, and a degree of external dissimilarity measure of this value with regard to all other values of the attribute for objects of other categories on the other hand. The most typical value of the attribute for a category corresponds to the maximum degree of typicality. A prototype is then an abstract object characterized by the most typical value of each attribute.

The variety of aggregation operators used to combine internal resemblance and external dissimilarity provides a flexible definition of a prototype balancing the relative importance of the common points of the category members and their distinctive properties as opposed to other categories.

It is also possible to consider objects globally and not attribute by attribute. The same principle leads to the computation of the internal resemblance of an object with respect to other members of the same category, and the external dissimilarity of this object with respect to members of other categories [58].

The particular case of numerical data has been considered in [58][59]. Exceptions have been less studied in cognitive science. They can be regarded as elements with a small typicality degree in all categories. Taking them into account in the construction of categories is a difficult problem and they are often considered as a nuisance in clustering methods. Some real-world problems require them to be considered as very informative elements and clustering methods have been settled to take them into account [60].

5 Examples of Utilization in Real Word Applications

Similarities and prototypes have been extensively used in fuzzy data mining. Similarity measures are obviously present in most of the systems based on fuzzy learning, fuzzy rule-based systems or fuzzy database management, for the evaluation of the degree of matching between a reference (attribute value in a decision tree, a rule, a query…) and all possible instances, examples or answers. We focus on systems in which the management of similarity is more complex.

To illustrate the above-mentioned use of similarities and dissimilarities, we will restrict this section to real-world applications that have been tackled in our research team [61]. We will distinguish mining in large amounts in data from information retrieval.

5.1 Image Interpretation

An example of environment where prototypes have been used to represent imprecise knowledge usually managed by medical doctors is the identification of microcalcifications [46][62]. Prototypes have been used to establish a link between linguistic criteria used by specialists to describe spots in mammographical images, for instance "round" or "small", and technical attributes, such as surface, convexity or elongation of the objects. Prototypes of "round" spots can for instance be described by fuzzy values of attributes [62], with simplified interpretations of the form "circularity is approximately 5 or 6, no more than 10" and "circumference is either approximately 6 or approximately 12" and…"

The resemblance measure that has proven to be the most successful with regard to the tests of classification is the following:

$$S(A, B) = \frac{2}{\pi} \arctan\left(2 * M(A \cap B) * \sup(A \cap B) / M(A \cup B)\right) \quad (12)$$

where M is the surface under the membership function.

5.2 Defect Forecasting

Fuzzy association rules can be chosen to extract knowledge from large databases. We have used this approach to forecast defects in gas pipelines. In-line inspections by means of smart pigs are used by gas operators, but they are not satisfying with regard to the exact dimensioning of the defects and real defects do not exactly correspond to those forecasted by the smart pig. We have used association rules to compare forecasted defects and real ones [23]. Fuzzy descriptions have been introduced to

represent linguistic expertise. A measure of satisfiability has been chosen to compare observed data and fuzzy descriptions, as follows:

$$S(A, B) = \frac{2}{\pi} \arctan\left(\frac{M(A \cap B)}{M(B \Theta A)}\right) \tag{13}$$

5.3 Risk Rating

Risk prediction and analysis is a complex topic, subject to imprecision and uncertainty in data and to linguistic expert knowledge. We have faced the problem of country risk ratings and a methodology to assess internal conflict risk has been proposed, with various components [63]. In the case of dynamic early warning, scenarios have been elaborated, taking into account temporal constraints, on the basis of human expertise [64]. For a given piece of information regarding a country, a satisfiability measure is used to compare its description with a scenario and the obtained results provide hypotheses that will be confirmed or refuted in the future. In order to obtain automatically elements of the scenarios, prototypes have been constructed and the following resemblance and dissimilarity measures have been chosen for the quality of the obtained results:

$$S(A, B) = \frac{M(A \cap B)}{M(A \cup B)}, \tag{14}$$

$$D(A, B) = 1 - \exp\left(-\frac{M(A \Theta B) + M(B \Theta A)}{\Gamma}\right). \tag{15}$$

The interpretation of a prototype of the category "ethnic conflict" is for instance of the form "number of extended military aid approximately between 3 and 4, and number of ultimatum approximately between 4 and 5, and …" The obtained fuzzy descriptions provide a prototypical vision of the category.

5.4 Web Usage Mining

Web usage mining requires recording and management of large amounts of web log files. A method has been conceived to select informative data and to construct prototypes of the activity of users on a website in order to provide a meaningful visualization of categories of users with similar navigation behavior [65]. Such a tool can help to improve the quality of a website through a better understanding of the expectations of prototypical users, or to provide an adaptive pedagogical support to learners according to the category they belong to, if the website is used in e-learning for instance.

In this case, the considered data are not fuzzy, thus the applied similarity measure does not belong to the framework described in the previous section. More precisely, the similarity must compare user web sessions, described in terms of the accessed web pages: two sessions are then considered as similar if they access similar pages. The similarity between web pages is based on their url addresses, and not on their content, so as to avoid an extraction and indexation step with high computational cost.

This approach relies on the assumption that the structure of the web site directory reflects its content. The similarity between two urls $S_{url}(u_1,u_2)$ is computed as the weighted sum of the elements identical in both paths. The normalized similarity between two sessions, s_1 and s_2, is then defined as

$$S(s_1,s_2) = \frac{\tilde{S}(s_1,s_2)}{\sqrt{\tilde{S}(s_1,s_1)\tilde{S}(s_2,s_2)}} \text{ with } \tilde{S}(s_1,s_2) = \sum_{u_1 \in s_1} \sum_{u_2 \in s_2} S_{url}(u_1,u_2) \qquad (16)$$

It is to be noticed that although the considered data are not fuzzy, the fuzzy similarity-based typicality framework can be used to identify the most typical user of each cluster, to characterize the identified clusters, i.e. the identified navigation behaviors.

5.5 Content-Based Image Retrieval

Image retrieval on large databases can be based on annotated documents or on a comparison of images on the basis of their content. An experimental platform has been proposed [66] for a retrieval of images similar to a given one considered as a reference. This image is automatically segmented into regions; attributes of regions such as colour or position in the image are taken into account. Colour histograms can be regarded as fuzzy sets of a universe of colors. The user chooses the regions of the reference image he wants to retrieve, to indicate the attributes he wants to consider and their importance in his query. Various measures of satisfiability were proposed and the result of the query was a list of images satisfying the query in a decreasing order of the satisfiability degree. This result was obviously dependent on the equivalence class of satisfiability measures [46], allowing the choice of a measure by the user to be restricted to those providing distinct orders.

6 Conclusion

Studies about similarity are countless and artificial intelligence takes advantage of studies in cognitive science in the construction of automated systems. In particular, the main streams of research on similarity and categorization in psychology have given rise to interesting foundations for developments in data mining. Because of the graded structure of natural categories and their variability, fuzzy set theory has been a privileged component of formalized versions of similarities and categories.

Our purpose has been to point out the richness of the concepts of similarity and category and some of their utilizations in fuzzy data mining. More directions remain to be explored.

Some of them have already been approached in artificial intelligence. Wittgenstein's concepts [5] in linguistics have been used by M. Sugeno and his colleagues as a semiotic base of an everyday language computing system, for instance in [67].

Let us just mention two other directions worth to develop in fuzzy data mining, dealing with dynamic similarities presented in Section 2.

In document retrieval, the choice of a similarity providing interesting answers to a user's queries is not simple. This choice is generally left to the expert. Determining the best measure of similarity for a user can also be done in an adaptive manner, on the basis of his interactions with the system. An example of such a method is based on a representation of retrieved images by means of Self Organizing Maps and a manual assignment of images to classes by the user [63]. The system learns such assignments to adapt its behaviour to the user's preferences, in an adaptive similarity management.

Another utilization of dynamic similarities concerns evolving categories. Methods of novelty detection, incremental classification methods or adaptive classification algorithms are based on the reorganization of classes or clusters according to the incoming data, in [69][70][71]for instance.

References

1. Rosch, E.: Principles of categorization. In: Rosch, E., Lloyd, B. (eds.) Cognition and Categorization. Lawrence Erlbaum, Mahwah (1978)
2. Rosch, E., Mervis, C.: Family resemblance: studies of the internal structure of categories. Cognitive psychology 7, 573–605 (1975)
3. Tijus, C.: Introduction à la psychologie cognitive. Nathan Université (2001)
4. Hampton, J.A.: The role of similarity in natural categorization. In: Hahn, U., Ramscar, M. (eds.) Similarity and Categorization, pp. 13–28. Oxford University Press, Oxford (2001)
5. Wittgenstein, L.: Philosophical Investigations. Blackwell Publishing, Malden (1953/2001)
6. Kleiber, G.: Prototype et prototypes. In: Sémantique et cognition. CNRS, Paris (1991)
7. Barsalou, L.W.: Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure. Journal of Experimental Psychology: Learning, Memory and Cognition 11, 629–654 (1985)
8. Nosofsky, R.M.: Similarity, frequency, and category representations. Journal of Experimental Psychology: Learning, Memory, and Cognition 14(1), 54–65 (1988)
9. Poitrenaud, S., Richard, J.-F., Tijus, C.: Properties, categories and categorization. Thinking and reasoning 11(2), 151–208 (2005)
10. Posner, M.I., Keele, S.W.: On the genesis of abstract ideas. Journal of Experimental Psychology 77, 353–363 (1968)
11. Barsalou, L.W.: The instability of graded structure: Implications for the nature of concepts. In: Neisser, U. (ed.) Concepts and conceptual development: Ecological and intellectual factors in categorization, pp. 101–140. Cambridge University Press, Cambridge (1987)
12. Tversky, A.: Features of similarity. Psychological Rev. 84(4), 327–352 (1977)
13. Hahn, U., Ramscar, M.: Introduction: similarity and categorization. In: Hahn, U., Ramscar, M. (eds.) Similarity and categorization, pp. 1–11. Oxford University Press, Oxford (2001)
14. Medin, D.L., Goldstone, R.L., Gentner, D.: Respects for similarity. Psychological Review 100(2), 254–278 (1993)
15. Keane, M.T., Smyth, B., O'Sullivan, F.: Dynamic similarity: A processing perspective on similarity (2001)
16. Markman, A.B., Gentner, D.: Structural alignment during similarity comparisons. Cognitive Psychology 25, 431–467 (1993)
17. Rissland, E.: AI and similarity, IEEE Int. Systems 21, 39–49 (2006)
18. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37–54 (1996)

19. Delavallade, T., Dang, T.H.: Using Entropy to Impute Missing Data in a Classification Task. In: IEEE International Conference on Fuzzy Systems, London, pp. 1–6 (2007)
20. Timm, H., Döring, C., Kruse, R.: Differentiated treatment of missing values in fuzzy clustering. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 354–361. Springer, Heidelberg (2003)
21. Song, Q., Shepperd, M.: A new imputation method for small software project data sets. Journal of Systems and Software 80(1), 51–62 (2007)
22. Setnes, M., Babuska, R., Kaymak, U., van Nauta Lemke, H.R.: Similarity measures in fuzzy rule base simplification. IEEE Transactions on Systems, Man, and Cybernetics, Part B 28(3), 376–386 (1998)
23. Pichlova, M., Bouchon-Meunier, B. : Using fuzzy association rules for defect forecasting in pipelines. Rencontres Francophones sur la Logique Floue et ses Applications, Nantes, Cépaduès-Editions, 305–312 (2004)
24. Marsala, C.: Fuzzy partitioning methods. In: Pedrycz, W. (ed.) Granular Computing: An Emerging Paradigm, pp. 163–186. Physica-Verlag GmbH, Heidelberg (2001)
25. Yuan, Y., Shaw, M.J.: Induction of Fuzzy Decision Trees. Fuzzy Sets and systems 69, 125–139 (1995)
26. Marsala, C., Bouchon-Meunier, B.: An adaptable system to construct fuzzy decision trees. In: Proceedings of the 18th International Conference of the North American Society, pp. 223–227 (1999)
27. Bouchon-Meunier, B., Marsala, C.: Linguistic modifiers and measures of similarity or resemblance. In: 9th IFSA World Congress, Vancouver, pp. 2195–2199 (2001)
28. Laurent, A., Marsala, C., Bouchon-Meunier, B.: Improvement of the Interpretability of Fuzzy Rule Based Systems: Quantifiers, Similarities and Aggregators. In: Davenport, J.H. (ed.) On the Integration of Algebraic Functions. LNCS (LNAI), pp. 102–123. Springer, Heidelberg (1981)
29. Hüllermeier, E.: Fuzzy-Methods in Machine Learning and Data Mining: Status and Prospects. Fuzzy Sets and Systems 156(3), 387–407 (2005)
30. Guillaume, S.: Designing Fuzzy Inference Systems from Data: An Interpretability-Oriented Review. IEEE Transactions on Fuzzy Systems 9(3), 426–444 (2001)
31. Zadeh, L.A.: Similarity relations and fuzzy ordering. Information Science, 177–200 (1971)
32. Valverde, L.: On the structure of F-indistinguishability operators. Fuzzy Sets and Systems 17, 313–328 (1985)
33. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles 44, 223–270 (1908)
34. Dice, L.R.: Measures of the amount of ecological association between species. Ecology 26, 297–302 (1945)
35. Ochiai, A.: Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bulletin of the Japanese Society for Science and Fisheries 22, 526–530 (1957)
36. Zwick, R., Carlstein, E., Budescu, D.V.: Measures of similarity among fuzzy concepts: A comparative analysis. International Journal of Approximate Reasoning 1, 221–242 (1987)
37. Chen, S., Yeh, M., Hsiao, P.: A comparison of similarity measures of fuzzy values. Fuzzy Sets Systems 72(1), 79–89 (1995)
38. Xuzhu, W., De Baets, B., Kerre, E.: A comparative study of similarity measures. Fuzzy Sets and Systems 73(2), 28, 259–268 (1995)
39. Jain, R., Murthy, S.N.J., Chen, P.L.-J., Chatterjee, S.: Similarity measures for image databases. In: IEEE International Conference on Fuzzy Systems, pp. 1247–1254 (1995)

40. Li, Y., Liu, J.-M., Li, J., Deng, W., Ye, C.-X., Wu, Z.-F.: The fuzzy similarity measures for content-based image retrieval. In: Proceedings of the International Conference on Machine Learning and Cybernetics, vol. 5, pp. 3224–3228 (2003)
41. Cross, V.V., Sudkamp, T.A.: Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications. Physica-Verlag (2002)
42. Dubois, D., Prade, H.: A unifying view of comparison indices in a fuzzy set-theoretic framework. In: Yager, R.R. (ed.) Fuzzy and possibility theory, pp. 3–13. Pergamon Press, Oxford (1982)
43. Shiina, K.: A fuzzy-set-theoretic feature model and its application to asymmetric data analysis. Japanese psychological research 30(3), 95–104 (1988)
44. Santini, S., Jain, R.: Similarity measures. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(9), 871–883 (1999)
45. Tolias, Y.A., Panas, S.M., Tsoukalas, L.H.: Generalized fuzzy indices for similarity matching. Fuzzy Sets Systems 120(2), 255–270 (2001)
46. Rifqi, M. : Mesures de comparaison, typicalité et classification d'objets flous: théorie et pratique. PhD thesis, Université Paris VI (1996)
47. Bouchon-Meunier, B., Rifqi, M., Bothorel, S.: Towards general measures of comparison of objects. Fuzzy Sets and Systems 84(2), 143–153 (1996)
48. Bouchon-Meunier, B., Rifqi, M.: OWA operators and an extension of the contrast model. In: Yager, R.R., Kacprzyk, J. (eds.) The Ordered Weighted Averaging Operators: Theory, Methodology, and Applications, pp. 29–35. Kluwer Academic Publishers, Dordrecht (1997)
49. Rifqi, M., Berger, V., Bouchon-Meunier, B.: Discrimination power of measures of comparison. Fuzzy Sets and Systems 110(2), 189–196 (2000)
50. Rifqi, M., Detyniecki, M., Bouchon-Meunier, B.: Discrimination power of measures of resemblance. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, Springer, Heidelberg (2003)
51. Omhover, J.-F., Detyniecki, M., Rifqi, M., Bouchon-Meunier, B.: Image Retrieval using Fuzzy Similarity: measure equivalence based on invariance in ranking. In: Proceedings of the IEEE International Conference on Fuzzy Systems, Budapest, Hungary, pp. 1367–1372 (2004)
52. Omhover, J.-F., Rifqi, M., Detyniecki, M.: Ranking Invariance based on Similarity Measures in Document Retrieval. In: Detyniecki, M., Jose, J.M., Nürnberger, A., van Rijsbergen, C.J. (eds.) AMR 2005. LNCS, vol. 3877, pp. 55–64. Springer, Heidelberg (2006)
53. Zadeh, L.A.: A note on prototype theory and fuzzy sets. Cognition 12, 291–297 (1982)
54. Friedman, M., Ming, M., Kandel, A.: On the theory of typicality. Int. Journ. of Uncertainty, Fuzziness and Knowledge-based Systems 3(2), 127–142 (1995)
55. Kacprzyk, J., Yager, R.: Linguistic summaries of data using fuzzy logic. Int. Journ. of General Systems 30, 133–154 (2001)
56. Lesot, M.-J., Rifqi, M., Bouchon-Meunier, B.: Fuzzy prototypes: from a cognitive view to a machine learning principle. In: Bustince, H., Herrera, F., Montero, J. (eds.) Fuzzy Sets and Their Extensions: Representation, Aggregation and Models Studies. Springer, Heidelberg (2007)
57. Rifqi, M.: Constructing prototypes from large databases. In: Proc. International Conference IPMU 1996, Granada, pp. 301–306 (1996)
58. Lesot, M.-J., Mouillet, L., Bouchon-Meunier, B.: Fuzzy prototypes based on typicality degrees. In: Proc. of Fuzzy Days 04, Springer, Advances on Soft Computing, pp. 125–138. Dortmund, Allemagne (2006)

59. Lesot, M.-J.: 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)
60. Lesot, M.-J.: Typicality-based clustering. Int. Journal of Information Technology and Intelligent Computing 1(2), 279–292 (2006)
61. Bouchon-Meunier, B., Detyniecki, M., Lesot, M.-J., Marsala, C., Rifqi, M.: Real world fuzzy logic applications in data mining and information retrieval. In: Wang, P.P., Ruan, D., Kerre, E.E. (eds.) Fuzzy Logic - A Spectrum of Theoretical and Practical Issues, Studies in Fuzziness, pp. 219–247. Springer, Heidelberg (2007)
62. Rifqi, M., Bothorel, S., Bouchon-Meunier, B., Muller, S.: Similarity and prototype-based approach for classification of microcalcifications. Int. J. General Systems 29(4), 623–636 (2000)
63. Delavallade, T., Mouillet, L., Bouchon-Meunier, B., Collain, E.: Monitoring event flows and modelling scenarios for crisis prediction, application to ethnic conflicts forecasting. Int. J. of Uncertainty, Fuzziness and knowledge-based systems, 15, 83–110 (2007)
64. Mouillet, L., Bouchon-Meunier, B., Collain, E.: Automated identification of political conflicts with a scenario recognition technique. In: 10th International Conference IPMU, Perugia, Italy, vol. 3, pp. 1609–1616 (2004)
65. Labroche, N., Lesot, M.-J., Yaffi, L.: A new web usage mining and visualization tool. In: IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Patras, Greece, pp. 321–328 (2007)
66. Omhover, J.-F., Detyniecki, M.: STRICT: an Image Retrieval Platform for Queries Based on Regional Content. In: Enser, P.G.B., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A.F., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 473–482. Springer, Heidelberg (2004)
67. Kobayashi, I., Sugeno, M.: An approach to a dynamic system simulation based on human information processing. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(6), 611–633 (2002)
68. Detyniecki, M., Nürnberger, A.: Adaptive multimedia retrieval: from data to user interaction. In: Gabrys, B., Leiviska, K., Strackeljan, J. (eds.) Do Smart adaptive systems exist – Best practice for selection and combination of intelligent methods. Series on Studies on Fuzziness and Soft Computing, pp. 341–370. Springer, Heidelberg (2004)
69. Utgoff, P.E.: Incremental Induction of Decision Trees. In: Machine Learning, vol. 4, pp. 161–185 (1989)
70. Wang, T., Li, Z., Yan, Y., Chen, H.: An Incremental Fuzzy Decision Tree Classification Method for Mining Data Streams. In: Perner, P. (ed.) MLDM 2007. LNCS (LNAI), vol. 4571, pp. 91–103. Springer, Heidelberg (2007)
71. Prehn, H., Sommer, G.: An Adaptive Classification Algorithm Using Robust Incremental Clustering. In: 18th International Conference on Pattern Recognition, vol. 1, pp. 896–899 (2006)