Similarity and prototype based approach for classification of microcalcifications

M. Rifqi[a], S. Bothorel[b], B. Bouchon-Meunier[a], S. Muller[b]

[a] *LIP6 – Université Pierre et Marie Curie*
Case 169 – 4, Place Jussieu, 75252 Paris Cedex 05
{bouchon,rifqi}@laforia.ibp.fr

[b] *General Electric Medical Systems Europe*
283, route de la Minière, 78533 Buc Cedex
{bothorel,mullese}@gemse.fr

Abstract. Our aim is to show the utility of a formal framework of measures of comparison, especially for a similarity based classification. We present both theoretical and practical arguments and we apply this approach to a real problem.

1 Introduction

An important number of classification methods are based on the comparisons of objects : the k-nearest neighbors (k-NN) method or instance based learning [8], [1], clustering methods like [10], [11],... The measure used to compare objects is often a distance. But, more and more, a similarity or a dissimilarity measure is chosen.

It is not easy to choose an appropriate measure. The choice is linked to the problem of the characterization of relevant properties for a classification task.

In this paper, we use the formalization and the framework introduced in [6] to deal with measures of comparison. We test this framework in a challenging classification problem: the classification of microcalcifications in mammography. Furthermore, we test a classification method based on fuzzy prototypes proposed in [13].

2 Similarity based-classification

2.1 The choice of a measure of similarity

A similarity based-classification method has to solve the problem of the choice of a measure of similarity or, more generally, a family of measures of comparison. In [6], we propose to formalize a measure of comparison between two fuzzy sets as a function of the common features and the distinctive features.

Formally, for any set Ω of elements, let $F(\Omega)$ denote the set of fuzzy subsets of Ω, f_A the membership function of any description A in $F(\Omega)$ and for any fuzzy set measure M, the definition is:

Definition 1 *An* M*-measure of comparison on* Ω *is a mapping* $S : F(\Omega) \times F(\Omega) \rightarrow [0,1]$ *such that* $S(A,B) = F_S(M(A \cap B), M(B - A), M(A - B))$, *for a given mapping* $F_S : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow [0,1]$ *and a fuzzy set measure* M *on* $F(\Omega)$.

An M-measure of comparison can either evaluate the likeliness of two descriptions (it is called an M-measure of similitude), or their differences (it is then called an M-measure of dissimilarity).

We consider the following properties for F_S:

- symmetry: $F_S(u,v,w) = F_S(u,w,v)$
- reflexivity: $F_S(u,0,0) = 1$
- containment: $F_S(u,0,w) = 1$ whatever $u \neq 0$ and w may be.
- exclusiveness: $F_S(0,v,w) = 0$ whatever v and w may be.
- minimality: $F_S(u,0,0) = 0$

The measure of dissimilarity is not defined as the dual of a measure of similitude, but it has specific properties.

Definition 2 *An* M*-measure of dissimilarity* S *on* Ω *is an* M*-measure of comparison satisfying the minimality property and such that* $F_S(u,v,w)$ *is independent of* u *and non decreasing in* v *and* w.

A definite symmetrical M-measure of dissimilarity satisfying the triangular inequality is a distance.

Definition 3 *An* M*-measure of similitude* S *on* Ω *is an* M*-measure of comparison* S *such that* $F_S(u,v,w)$ *is non decreasing in* u, *non increasing in* v *and* w.

The relation given by Tversky [20] and generalized to fuzzy sets [18], [6]: $S(A,B) = f(A \cap B)/(f(A \cap B) + \alpha f(A - B) + \beta f(B - A))$ $\alpha, \beta \geq 0$ is an f-measure of similitude if f is a fuzzy set measure.

In order to classify the different measures more subtly, we distinguish three types of M-measures of similitude: satisfiability, inclusion and resemblance.

A measure of *resemblance* 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.

Definition 4 *An* M-measure of resemblance *on* Ω *is an* M*-measure of similitude* S *which satisfies the reflexivity and the symmetry properties.*

M-measures of resemblance S which satisfy an additional property of T-transitivity are extensions of indistinguishability relations [19], [21] to fuzzy sets. In the case where T is the minimum, we obtain extensions of measures of similarity.

Measures of resemblance are appropriate for a case-based reasoning or an instance based-learning. In clustering methods, distances can be replaced by a measure of resemblance. More generally, similarity-based classification methods has to use resemblance measure as soon as all objects have the same level of generality.

A measure of *satisfiability* 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 the reference.

Definition 5 *An* M-measure of satisfiability S *on* Ω *is an* M*-measure of similitude* S *satisfying the containment and the exclusiveness properties and such that* $F_S(u, v, w)$ *is independent of* w*.*

Analogy relations [7]: $S(A,B) = \inf_x \min(1 - f_B(x) + f_A(x), 1)$, and fuzzy similitude [3]: $S(A,B) = 1 - \sup_{f_A(x)=0} f_B(x)$ are particular M-measure of satisfiability.

Measures of satisfiability have been proven [6] to be compatible with the contrast model introduced by Tversky, satisfying major properties such as matching, monotonicity, independence, solvability [20].

Measures of satisfiability are appropriate for rule base systems. For example, in [4] or in [2] objects are classified by means of a decision tree. In a decision tree, a node represents a test on the chosen attribute during the learning stage; each edge of this node is associated with a value of the attribute. The classification of a new object comes to find consecutive edges from the root to the leaves. In [4] and in [2], the comparison between the value of an attribute of the new example with test-values associated with each edge is realized by means of a measure of satisfiability.

A measure of *inclusion* also concerns a situation with a reference object and measures if the points common to A and B are important with regard to A.

Definition 6 *An* M-measure of inclusion S *on* Ω *is an* M*-measure of similitude satisfying the reflexivity and the exclusiveness properties such that* $F_S(u, v, w)$ *is independent of* v*.*

As an example of M-measure of inclusion, we can find the degree introduced by Sanchez [17]: $S(A,B) = |A \cap B|/|A| = M(A \cap B)/(M(A \cap B) + M(A - B))$ with the sigma-count as M.

Measures of inclusion are appropriate to database management. For example, [12] et [16] use measures of inclusion for their database system in order to compare different classes.

2.2 Description of the classification method

Before the very step of classification, objects has to be compared in order to construct a prototype for each class.

2.2.1 Construction of a fuzzy prototype

According to E. Rosch [14], objects do not represent all in a same manner the category they belong. They are spread along a scale of typicality. According to Rosch and Mervis [15] :

> *[..] categories tend to become defined in terms of prototypes or prototypical instances that contain the attributes most representative of items inside and least representative of items outside the category.*(p.30)

Then, the notion of prototype is linked to the notion of typicality. Zadeh [22] has also emphasized this aspect: the typicality is a question of degree and it implies that the concept of prototype is a fuzzy concept.

In our method, we need to determine the typicality of each value appearing in a learning database in order to construct a fuzzy prototype.

Degree of typicality We consider that the degree of typicality of an object depends positively on its total resemblance to others objects of its class (internal resemblance) and on its total dissimilarity to objects of other classes (external dissimilarity). The term "resemblance" refers here to measures of resemblance. Indeed, objects are compared two by two in order to determine their total resemblance to others objects of its class. This situation of comparison corresponds to a situation where objects are considered to have the same level of generality. No value can be taken as a reference. So, this situation needs a measure of resemblance as a measure of similitude.

Let X be a set of objects. We suppose that there exists a partition given on X composed by crisp classes $\mathcal{C}_j$. The typicality of the value v of an

attribute A of an object O of the class $\mathcal{C}_i$ is computed as follows :

Step 1. Compute the resemblance $r(v, v_j)$ between v and the value v_j of the attribute A for any example of the same class $\mathcal{C}_i$. The global resemblance $R(v)$ relative to the set of values of A present in examples, is obtained in aggregating the degrees $r(v, v_j)$ computed as above described.

Step 2. Compute the dissimilarity $d(v, v_j)$ between v and the value v_j of the attribute A for any example of class $\mathcal{C}_k$ different from $\mathcal{C}_i$. The total dissimilarity $D(v)$ relative to the set of values of A present in examples, is obtained in aggregating the degrees $d(v, v_j)$ computed as above described.

Step 3. The aggregation of this two values, $R(v)$ et $D(v)$, gives the typicality $T(v)$ of v, according to the attribute A, for the class $\mathcal{C}_i$.

Fuzzy prototype Degrees of typicality participate in the construction of a fuzzy prototype of a given class. For an attribute A, the degree of typicality of each value of A is computed for each class. Then, the fuzzy prototype is composed by the most typical value(s) for each attribute of a considered class. That is to say that a fuzzy prototype is a virtual object described by means of the same attributes that those pertaining to the learning database. The values taken by the fuzzy prototype are the most typical.

A prototype, as said Zadeh [22], is not a unique object or a group of objects. It is more a fuzzy schema enabling to generate a set of objects because of the synthesized information it contains. For Desclés [9], this generation of objects is possible by means of successive determination of the prototype.

The prototype is intrinsically interesting because of its power of description. This power can be used for a classification process.

2.2.2 Classification

A new object with an unknown class is classified thanks to a comparison with each prototype of each class. Indeed, a prototype can be considered as a rule describing a class [5]. For example, the prototype of the class "scientific student" might be: `very good` in mathematics, physics, `medium` in literature and foreign language. In other words, `if` mathematics = `very good`, `and` physics = `very good`, `and` literature = `medium and` foreign language = `medium then` class = scientific student.

The classification process is based on the question: *does the new object satisfy a prototype?* This question entails the use of a *measure of satisfiability* for each comparison. The computed degrees of satisfiability are aggregated in order to obtain a total degree of satisfiability of a new object for a prototype.

3 Application

3.1 Description of the problem

One woman in 8 in the United States and one woman in 10 in Europe will have a breast cancer during her life. Nowadays, mammography is the primary diagnostic procedure for the early detection of breast cancer. Microcalcification clusters are an important element in the detection of breast cancer. This kind of finding is the direct expression of pathologies which may be benign or malignant. The objective is to help radiologists to increase their sensitivity in both detection and characterization task.

The description of microcalcifications is not an easy task, even for an expert. If some of them are easy to detect and to identify, some others are more ambiguous. The texture of the image, the small size of objects to be detected (less than one millimeter), the various aspects they have, the radiological noise, are parameters which impact the detection and the characterization tasks (Figure 1).

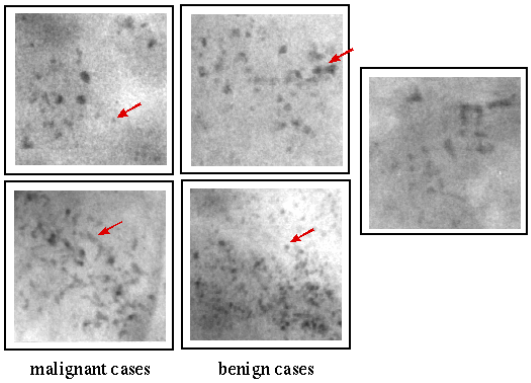

Figure 1 – Imprecision and uncertainty of the contours of microcalcifications

More generally, mammographic images present two kinds of ambiguity: *imprecision* and *uncertainty*. The *imprecision* on the contour of an object comes from the fuzzy aspect of the borders: the expert can define approximately the contour but certainly not with a high spatial precision. The *uncertainty* comes from the microcalcification superimpositions: because objects are built from the superimpositions of several 3D structures on a single image, we may have a doubt about the contour position.

Following the human reasoning, the classical image processing chain for image interpretation has three steps :

- Detection / segmentation of objects
- Object characterization / extraction and transformation of the image information
- Classification / diagnostic

An algorithm, proposed in [2], has been developed in order to characterize microcalcification clusters, following these steps.

3.2 Our goal

The challenging problem is to design an algorithm to recognize a malignant cluster of microcalcifications in order to help radiologists to detect automatically the breast cancer. Identification and contouring of the microcalcifications are performed before analyzing clusters. This first stage is illustrated in Figure 2.

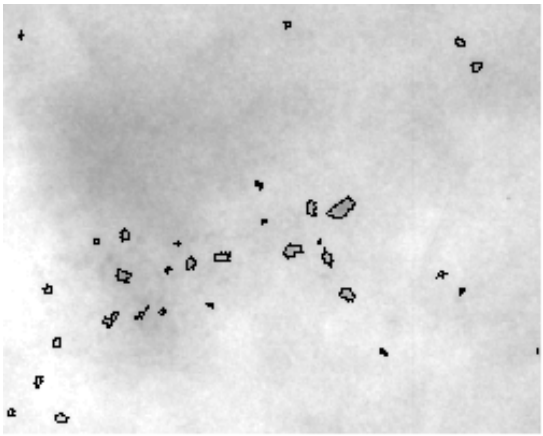

Figure 2 – Contours of microcalcifications

In this paper, we focus on the microcalcifications classification. The two first steps of image interpretation are considered to be solved. Then, we dispose of the microcalcifications descriptions, classified in round or not, small or not, long-shaped or not. A learning database and a test database are supposed to be given for each characterization (round or not, small or not, long-shaped or not). Our goals are:

- to describe each class (round or not, small or not, long-shaped or not) by a fuzzy prototype
- to classify an unknown microcalcification by comparing it with each fuzzy prototype.

The sizes of learning and test databases are given in Tab 1.

	Learning database	*Test database*
Round	28	39
Not round	66	69
Long	42	43
Not long	100	93
Small	107	118
Not small	43	41

Tab 1 – Size of learning and test database.

Each microcalcification is described by means of 7 fuzzy attributes. These 7 attributes enable to describe more precisely:

- the contrast (1 attribute)
- the shape (3 attributes) : elongation, compacity1, compacity2.
- the dimension (2 attributes) : surface, perimeter.
- the volume (1 attribute)

Figure 3 gives an example of a microcalcification classified "small". Among seven attributes, only two are presented here: surface and compacity.

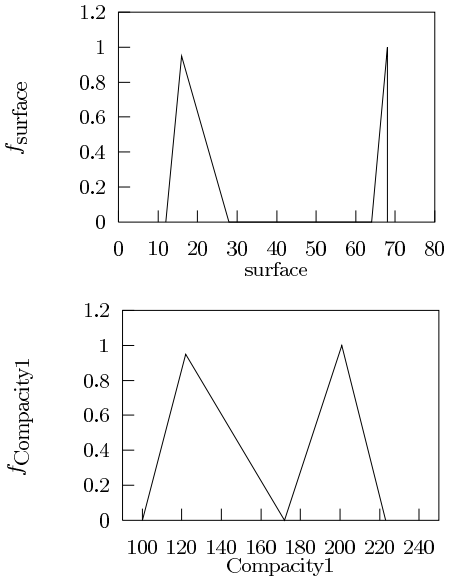

Figure 3 – Description of a microcalcification by means of fuzzy values.

We present here the best prototypes obtained for the class "round" and "not round" (Figure 4), regarding the rate of classification totalized. The results of our method and of k-NN method of classification for different classes are given in Tab 2.

	Our method	k-NN
Round/Not round	82.41	79.63
Long/Not long	80.88	73.53
Small/Not small	93.71	91.82

Tab 2 – Results of classification in percent of well-classified.

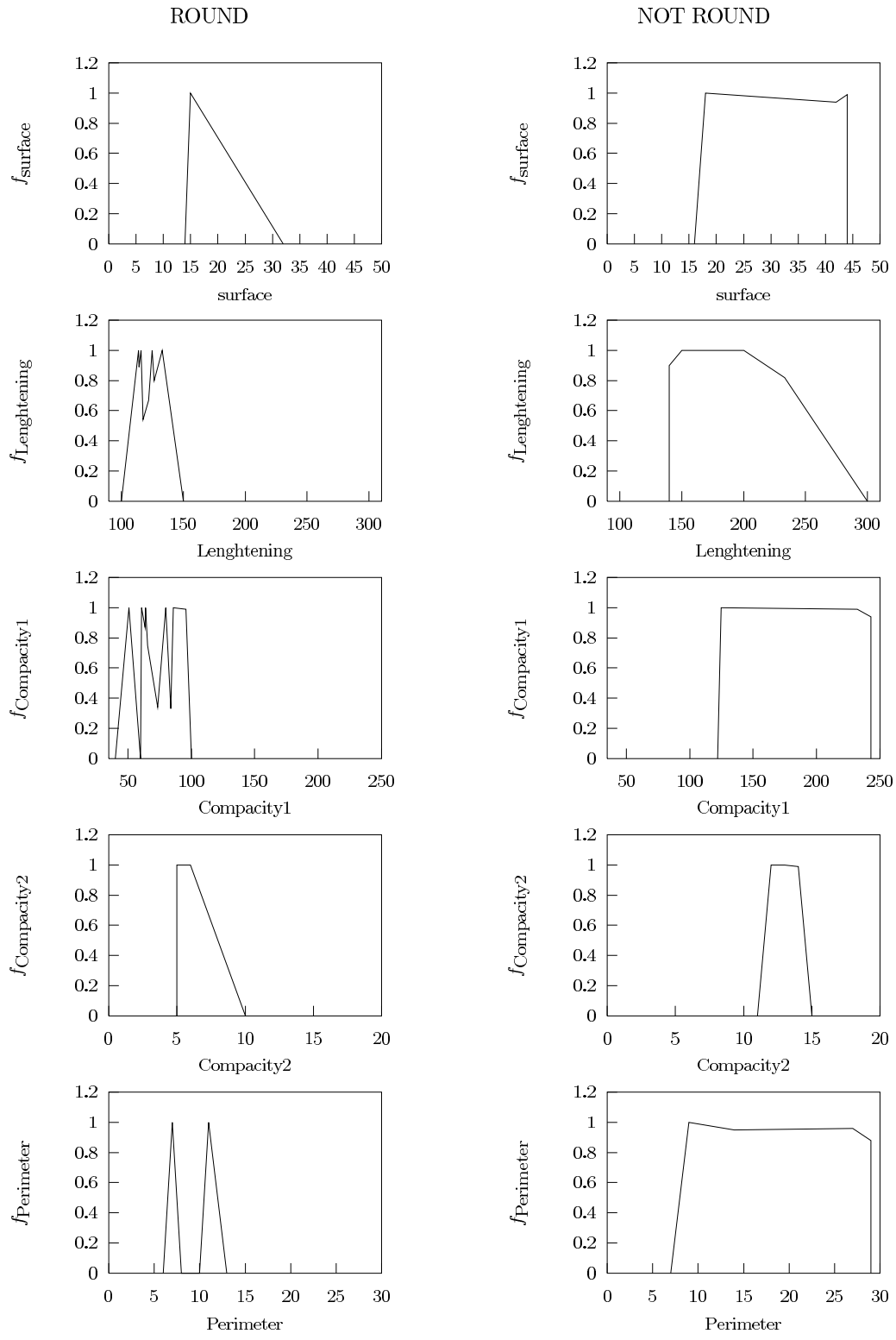

Figure 4 – Prototypes of the classes "round" and "not round"

4 Conclusion

The formal framework of comparison measures we have proposed has been tested with a real problem. This test has confirmed that this framework is performant. Furthermore, degrees of typicality based on comparison measures are effective for the construction of fuzzy prototypes. These prototypes are also effective for a classification problem.

Aknowledgment

We are grateful to Doctor Levy (Institut de Radiologie – Scanner Hoche, Paris – France) for providing us the original films used in this study.

References

1. D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.
2. S. Bothorel. *Analyse d'image par arbre de décision flou. Application à la classification semiologique des amas de microcalcifications.* PhD thesis, Université Paris 6, Décembre 1996.
3. B. Bouchon-Meunier. Fuzzy similitude and approximate reasoning. In P. P. Wang, editor, *Advances in Fuzzy Theory and Technology*, pages 161–166. Bookwrights Press, 1993.
4. B. Bouchon-Meunier, C. Marsala, and M. Ramdani. Learning from imperfect data. In H. Prade D. Dubois and R. R. Yager, editors, *Fuzzy Information Engineering: a Guided Tour of Applications*, pages 139–148. John Wileys and Sons, 1997.
5. B. Bouchon-Meunier, C. Marsala, and M. Rifqi. Comment classer des objets imparfaitement décrits ? In *XXV^{ème} colloque international de l'ARAE (Association Rhodanienne pour l'Avancement de l'Econométrie) sur les Structures économiques, l'économétrie et l'informatique*, Lyon, France, Mai 1996.
6. B. Bouchon-Meunier, M. Rifqi, and S. Bothorel. Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84(2):143–153, 1996.
7. B. Bouchon-Meunier and L. Valverde. Analogy relations and inference. In *Proceedings of 2nd IEEE International Conference on Fuzzy Systems*, pages 1140–1144, San Fransisco, 1993.
8. B. V. Dasarathy. *Nearest Neighbors (NN) Norms: NN pattern classification techniques.* IEEE Computer Society Press, 1990.
9. J.-P. Desclés. Implication entre concepts : la notion de typicalité. *Travaux de linguistique et de littérature, Centre de philologie et de littératures romanes de l'Université de Strasbourg*, XXIV(1):179–202, 1986.
10. E. Diday. Introduction à la méthode des nuées dynamiques. In *Analyse des données*, volume I, pages 121–132. A. P. M. E. P (Association des Professeurs de Mathématiques de l'Enseignement Public), 1980.
11. E. Diday, J. Lemaire, J. Pouget, and F. Testu. *Eléments d'analyse de données.* Dunod, 1982.
12. M-N Omri. *Système interactif flou d'aide à l'utilisation de dispositifs techniques : le sytème SIFADE.* PhD thesis, Université Pierre et Marie Curie, Paris, France, 1994.
13. M. Rifqi. Constructing prototypes from large databases. In *IPMU'96*, pages 301–306, Granada, 1996.
14. E. Rosch. Principles of categorization. In E. Rosch and B. B. Lloyd, editors, *Cognition and categorization*, pages 27–48. Hillsdale, N. J. : Laurence Erlbaum Associates, 1978.
15. E. Rosch and C. B. Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.
16. J-P Rossazza. *Utilisation de hiérarchie de classes floues pour la représentation des connaissances imprécises et sujettes à exception : le système SORCIER.* PhD thesis, Université Paul Sabatier, Toulouse, France, 1990.
17. E. Sanchez. Inverses of fuzzy relations. Applications to possibility distributions and medical diagnosis. *Fuzzy Sets and Systems*, 2(1):75–86, 1979.
18. K. Shiina. A fuzzy-set-theoretic feature model and its application to asymmetric data analysis. *Japanese psychological research*, 30(3):95–104, 1988.
19. E. Trillas and L. Valverde. On implication and indistinguishability in the setting of fuzzy logic. In J. Kacprzyk and R. R. Yager, editors, *Management Decision Support Systems Using Fuzzy Sets and Possibility Theory.* Verlag TUV, Rheinland, 1984.
20. A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.
21. L. Valverde. On the structure of t-indistinguishability operators. *Fuzzy Sets and Systems*, 17:313–328, 1985.
22. L. A. Zadeh. A note on prototype theory and fuzzy sets. *Cognition*, 12:291–297, 1982.