TYPICALITY-BASED CLUSTERING

Marie-Jeanne LESOT[1]

[1]*Department of Knowledge Processing and Language Engineering,*
Otto-von-Guericke Universität Magdeburg, Germany
E-mail: lesot@iws.cs.uni-magdeburg.de

Abstract

Typicality degrees are defined to build prototypes that characterise data subcategories, taking into account both the common points of the category members and their distinctive features as compared to other categories. In this paper, these principles are extended to the unsupervised learning framework, leading to a clustering algorithm robust to outliers that avoids overlapping areas between clusters and builds subgroups that are indeed both compact and separable. It does not require to use a Euclidean distance, which makes it possible to identify non-convex clusters.

Keywords: clustering, typicality degrees, outliers, non-convex clusters

1   Introduction

Clustering is an unsupervised learning task that aims at decomposing a data set into compact and distinct subgroups: compactness justifies the data grouping, as cluster members indeed resemble one another; the fact that clusters are distinct, or their separability, justifies the individual existence of each group, as merging them would lead to lose the compactness property. Clustering provides a simplified representation of the data set, highlighting its structure and helping the user to better apprehend it.

In this paper, a clustering algorithm based on typicality degrees is proposed. The latter were defined in a prototype construction procedure [8, 6] to measure data point representativeness: all members of a category do not have the same status, some are better examples of the category, they are more characteristic, or more representative for it. For instance in the case of the mammal category, a dog can be considered as more typical than a platypus or a bat.

According to cognitive science results [9, 10], the typicality of a point for its category depends on two complementary notions, called internal resemblance and external dissimilarity: the former is defined as the point resemblance to the other members of the category, the latter as its dissimilarity from members of the category. This can be illustrated with the previous mammal example: a platypus is atypical mainly because it does not resemble enough other mammals, whereas a bat is atypical mainly because it is not different enough from the members of the bird category. Prototypes based on such typicality degrees then lead to category representatives that highlight both the common features of the category members and their discriminative features as opposed to other categories, underlining their specificity.

These principles were implemented to automatically extract relevant characteristic representations of data categories in a supervised learning framework [8, 6]. In this paper, they are adapted to unsupervised learning, to perform clustering i.e. identify relevant subgroups in a data set: the internal resemblance and external dissimilarity on which typicality relies can be matched with the compactness and separability properties that are desired from clusters. Indeed, a compact cluster means that all cluster members have a high internal resemblance; well-separated clusters imply all points have a high external dissimilarity. Thus a cluster decomposition has a high quality if all points have a high typicality degree for the cluster they are assigned to. Therefore, we propose a clustering algorithm that aims at maximising the typicality degrees. As detailed in the following, the algorithm is robust to outliers, avoids overlapping areas between clusters and can be used with any comparison measures, making it possible to detect non-convex clusters.

The paper is organised as follows: Section 2 recalls the typicality degree formalisation and Section 3 presents the proposed extension to unsupervised learning. Section 4 illustrates the results obtained with an artificial data set and compares the proposed method to other clustering algorithms. Lastly Section 5 illustrates the algorithm application to detect non-convex clusters.

2 Typicality Degrees

A prototype is an element chosen to represent, characterise and summarise a data set. Rifqi [8] proposed a construction method, later extended in [6], relying on the typicality notion defined by Rosch [9]: according to the latter, the typicality of a point for the category it belongs to depends on its resemblance to the other members of the category (internal resemblance), and on its dissimilarity to members of other categories (external dissimilarity). The pro-

totype derived from such typicality degrees then underlines both the common and discriminative features of the category members.

Given a data set $X = \{x_i, i = 1..n\}$ with points belonging to several categories, a category C, and a point $x \in C$, the method first computes [8]

$$\text{the internal resemblance} \quad R(x, C) = avg(\rho(x, y), y \in C) \tag{1}$$

$$\text{the external dissimilarity} \quad D(x, C) = avg(\delta(x, y), y \notin C) \tag{2}$$

where ρ (resp. δ) is a resemblance (resp. dissimilarity) measure, i.e. a function that gives a value in the interval $[0, 1]$ measuring the extent to which the two points are similar (resp. different). One can cite as example normalised distances or normalised nonlinear transformations of the distance, for instance through the Cauchy function (see eq. (4)) or through a Gaussian function [1, 4]. Equation (1) defines the internal resemblance as the average resemblance to other members of the group, Equation (2) defines the external dissimilarity as the average dissimilarity to members of other categories.

Typicality degrees are then defined as the aggregation of the internal resemblance and external dissimilarity

$$T(x, C) = \varphi(R(x, C), D(x, C)) \tag{3}$$

where φ is an aggregation operator. Many possibilities can be considered, leading to various semantics for the typicality degrees [6]: conjunctive operators, such as the min, allow as typical points only those with both high internal resemblance and external dissimilarity. Disjunctive operators, such as the max, are less severe; they lead to a double semantic for typicality: two kinds of points can have high typicality degrees, namely those with high internal resemblance and those with high external dissimilarity. This can lead to non-convex typicality distribution. Trade-off operators, such as the weighted mean, make it possible to rule the relative importance of R and D and offer a compensation property: the decrease of one criterion can be compensated for by the increase of the other one. Lastly variable behaviour operators, such as the symmetric sum [2], offer a full reinforcement property: if both R and D are high, they reinforce each other to give an even higher typicality degree; likewise, if both are small, they penalise each other to give an even smaller typicality, in between, they offer a compensation property. (See [6] for a more complete discussion on aggregation operators for typicality computation.)

The prototype is then computed as the aggregation of the most typical data points: denoting τ a user-defined threshold and ψ an aggregation operator, it is defined as $p_C = \psi(\{x/T(x, C) > \tau\})$. In the case of crisp data, ψ can be a weighted mean, or a more complex operator that aggregates crisp values into a fuzzy set, so as to model the intrinsic imprecise property of prototypes [6].

3 Clustering Algorithm Based on Typicality Degrees

In this section, an extension of the previous typicality degrees to unsupervised learning is proposed, to identify relevant clusters in a data set. As indicated in the introduction, the underlying idea is to exploit the matching between internal resemblance and cluster compactness, as well as the matching between external dissimilarity and cluster separability. More precisely, the aim is to determine a data decomposition that maximises, for each data point, its typicality degree for the cluster it is assigned to.

To that aim, as summarised in table 1 and detailed in the following, the algorithm alternates two steps: (i) given a candidate partition of the data, typicality degrees are computed; (ii) given typicality degrees, a partition is computed, so that each point becomes more typical of the cluster it is assigned to.

Two advantages are expected from this approach, namely the exclusion of outliers and the avoidance of cluster overlapping areas: outliers have low internal resemblance whichever cluster they are assigned to, and thus should have low typicality degrees. Points locating in cluster overlapping areas are not distinct from points in other clusters and thus should also have low typicality degrees. This should lead to clusters that are both compact and separable.

3.1 Typicality Step

The first step of the algorithm consists in computing typicality degrees with respect to a candidate data partition. Contrary to the supervised case, typicality degrees are not computed only with respect to the cluster the point is assigned to, but with respect to all clusters. Indeed, the current cluster estimation is to be questioned and different assignments must be considered.

The candidate partition is only used to determine which points must be taken into account for internal resemblance and external dissimilarity: for a given point, when its typicality with respect to cluster C is computed, internal resemblance is based on points assigned to cluster C according to the candidate partition; likewise, external dissimilarity is computed with respect to points assigned to other clusters according to the candidate partition.

3.1.1 Comparison Measure Choice

As regards the resemblance and dissimilarity measures, involved in step 1a and 1b of the algorithm summary given in table 1, we consider Cauchy functions, as used in the possibilistic c-means (PCM) [3]: denoting d is the

Table 1. Proposed typicality-based clustering algorithm

Notations: $X = \{x_i, i = 1..n\}$ the data set, c the desired number of clusters, ρ and δ a resemblance and dissimilarity measure respectively, φ an aggregation operator

Initialisation: Apply a few steps of FCM and assign points according to their maximal membership degrees

Loop: while assignment evolves, alternate

1. Typicality step: for each point $x \in X$ and each cluster C_r, $r = 1..c$

   (a) Compute the internal resemblance $R(x, C_r) = avg(\rho(x, y), y \in C_r)$

   (b) Compute the external dissimilarity $D(x, C_r) = avg(\delta(x, y), y \notin C_r)$

   (c) Compute the typicality degree $T(x, C_r) = \varphi(R(x, C_r), D(x, C_r))$

2. Assignment step: for each point $x \in X$

   (a) if x is typical for no cluster, i.e. $\max_r T(x, C_r) < 0.1$, assign x to a fictitious cluster, C_0

   (b) else if x typicality is not clear, i.e. $T_1(x) - T_2(x) < 0.02$, where $T_i(x)$ is the i-th biggest value of $T(x, C_r)$, $r = 1..c$, assign x to the fictitious cluster C_0

   (c) else assign x according to the maximal typicality degree, i.e. to C_r where $r = \arg\max_s T(x, C_s)$.

Euclidean distance, γ_R and γ_D user-defined parameters corresponding to reference distances, they are defined as

$$\rho(x, y) = \frac{1}{1 + \left(\frac{d(x,y)}{\gamma_R}\right)^2} \qquad \delta(x, y) = 1 - \frac{1}{1 + \left(\frac{d(x,y)}{\gamma_D}\right)^2} \qquad (4)$$

γ_R and γ_D respectively indicate the distance from which the resemblance (resp. dissimilarity) is smaller (resp. higher) than 0.5. Now dissimilarity is used to compare points assigned to different clusters, and thus only considers inter-clusters distances, whereas resemblance is used to compare points belonging to the same cluster and has an intra-cluster meaning. Thus it is expected that dissimilarity applies to distances that are on average bigger than those involved in resemblance. This justifies the definition of different reference distances for resemblance and dissimilarity. For the dissimilarity, a reference distance is the data diameter: we choose γ_D so that dissimilarity is 0.9 for points at distance $diam(X)/2$. For resemblance, a local reference distance is defined, for each cluster independently, as the cluster radius, so that resemblance is 0.5 for points whose distance equals the cluster radius.

As clusters are not known, their radii are not known either. Therefore, two

definitions are used: in a preliminary step, clusters are initialised using a few iterations of fuzzy c-means (FCM). As in the PCM algorithm the following definition is used: for all $r = 1..c$, $\gamma_{Rr} = (\sum_i u_{ri} d(x_i, w_r))/(\sum_i u_{ri})$ where (w_r) is the position of the cluster r centre, and (u_{ri}) is the membership degree of point x_i for cluster r. Second, after having converged using these values, the obtained data partition is used to update the estimation of the cluster radii and to define γ_{Rr} as half the cluster r diameter.

3.1.2 Aggregation Operator Choice

As regards the aggregation operator φ involved in step 1c of the algorithm (see table 1), it must not give too much weight to the external dissimilarity: in the clustering case, one is not interested in discriminative prototypes whereas they can be justified for discrimination in supervised learning. For clustering, if external dissimilarity is given too much weight, outliers are considered as highly typical of any cluster and may disturb the clustering process.

Therefore, φ is first chosen to be a "severe" operator, such as the min. In a second step, when outliers have been excluded, one can be more tolerant, and use a variable behaviour aggregator, such as the symmetric sum, $\varphi(a, b) = ab/(ab + (1 - a)(1 - b))$ which offers a full reinforcement property [2].

3.2 Assignment Step

The second alternated step of the algorithm consists in modifying a data decomposition according to given typicality degrees, so that each data becomes more typical of the cluster it is assigned to. To that aim, as seems natural, points are assigned to the cluster they are most typical of

$$x_i \in C_r \Longleftrightarrow r = \arg\max_s T(x, C_s) \tag{5}$$

Two specific cases are handled differently: first points for which the maximal typicality degree is small (smaller than 0.1 in our experiments), i.e. points that are typical for no cluster. For such points indeed, the previous assignment scheme does not seem justified. These points correspond to outliers that should not be assigned to any cluster, but considered as aberrations. They are thus assigned to a fictitious cluster instead of a regular cluster.

Second, a special case is made of points for which the maximal typicality degree is not clear, i.e. the second biggest value is close to the biggest one (the difference is lower than 0.02 in our tests). Indeed, a tie-breaking strategy is necessary, the point assignment would not be justified: such points are also assigned to the fictitious cluster.
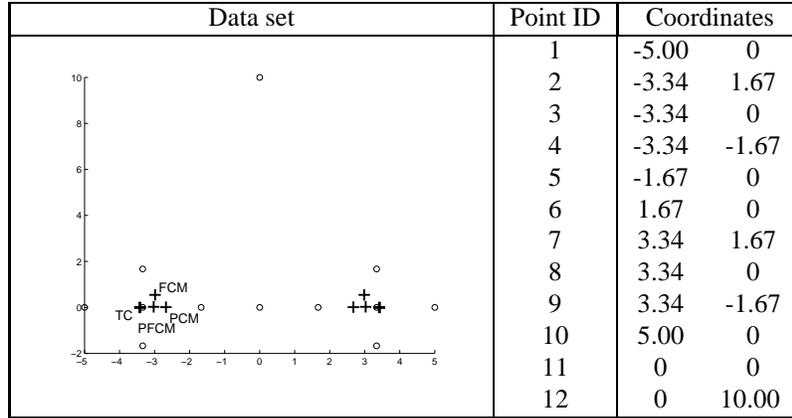
3.3 Overall Algorithm

Globally, the proposed algorithm first computes an initial partition of the data, through a few steps of the fuzzy c-means algorithm for instance, and deduces initial values for the cluster radii. It then applies the loop indicated in table 1, that consists in alternatively computing typicality degrees according to a partition and updating the partition according to typicality degrees, until convergence of the data partition. During this step, the comparison measures are the functions of eq. (4) with γ_R as in the PCM algorithm (see above) and $\varphi = \min$. The loop is then performed a second time, after updating the γ_R values and changing φ to the symmetric sum. In the performed tests, each convergence required only a very small number of iterations, less than 10.

4 Experimental Results

In this section, the proposed algorithm is compared to other clustering algorithms, on the data set used by Pal et al. [7], represented on figure 1. It contains two clusters of 5 points each and two specific points (numbered 11 and 12) located on the median between the two subgroups, at coordinates (0,0) and (0, 10) respectively, the second one corresponds to an outlier. We applied FCM, PCM, the Possibilistic Fuzzy c-means (PFCM) [7] and the proposed typicality-based algorithm, denoted TC (see caption of table 2 for the used parameters values). Figure 1 represents the position of the cluster centres obtained as weighted means of the data, using as weights the coefficients associated to each algorithm. The numerical values of both centre coordinates and weighting coefficients are indicated in table 2.

FCM results illustrate their sensitivity to outliers: point 12 attracts the centres that do not have ordinate 0 as expected, but ordinate 0.54. Indeed this point gets membership degree 0.5 for both clusters and thus has an important weight. It is as important as point 11, although the latter is less outlying. This is due to the fact that the FCM membership degrees involve relative distances, and do not decrease with the absolute distance to cluster centre. Due to a normalisation constraint, they are actually to be interpreted as sharing coefficients, that indicate the extent to which each point is shared between the clusters. As points 11 and 12 are both located on the median between the clusters, both are equally shared, FCM do no distinguish between them.

PCM [3] relax the FCM normalisation constraint, and the coefficients they rely on measure the absolute distance between data points and cluster centres. Therefore they take a value next to 0 for the outlier (point 12 is associated to coefficient 0.04), which thus does not attract the centre, leading to more

Data set	Point ID	Coordinates	
	1	-5.00	0
	2	-3.34	1.67
	3	-3.34	0
	4	-3.34	-1.67
	5	-1.67	0
	6	1.67	0
	7	3.34	1.67
	8	3.34	0
	9	3.34	-1.67
	10	5.00	0
	11	0	0
	12	0	10.00

Figure 1. Considered data set and cluster centre positions for some clustering algorithms. The numerical values of the coordinates of cluster centres are given in table 2.

satisfying results. PCM coefficients can be interpreted as measuring an internal resemblance, defined as the resemblance to the cluster centre; yet they do not take into account external dissimilarity. This is one of the reason why they suffer from a merging cluster problem (see [7, 12]): in some cases, clusters are coincident, whereas natural subgroups in the data are overlooked.

To answer this problem, Pal et al. [7] propose to combine FCM and PCM, i.e. to exploit both relative and absolute resemblances, respectively to perform assignment and to reduce the outlier influence. More precisely, denoting u_f and u_p FCM and PCM coefficients respectively, PFCM rely on coefficients defined as $au_f^m + bu_p^\eta$, with a, b, m and η user-defined parameters. Table 2 shows the PFCM centres are not attracted by the outlier, and are further apart from each other than the PCM centres: point 12 is associated to a small value, due to the PCM coefficient, that equals 0.0, whereas the FCM coefficient still equals 0.5. It is to be underlined that the PFCM coefficients cannot be interpreted as such; in particular, they are not normalised (see e.g. points 3 and 8).

At a theoretical level, the proposed typicality-based algorithm can be compared to PFCM insofar as it also combines two information types: it considers internal resemblance, that can be directly related to the PCM coefficients, and external dissimilarity, instead of FCM coefficients. These two components provide two points of view on the data, that can be regarded as more clearly complementary of each other than PCM and FCM coefficients. Furthermore, the aggregation is more flexible than that of PFCM, and lead to interpretable coefficients, contrary to PFCM.

At an experimental level, typicality degrees show similar coefficient values

Table 2. Coefficients and centre positions obtained using several clustering algorithms. FCM and PCM were applied with $c = 2$ and fuzzifier $m = 2$. PCM were initialised with FCM results and iterated twice, updating the normalising coefficients between the two steps. PFCM were used with the parameter values indicated in [7] leading to the closest to 0 centre ordinates ($a = 1$, $b = 1$, $m = 7$ and $\eta = 1.5$).

Point ID	FCM		PCM		PFCM		TC	
1	0.94	0.06	0.46	0.07	0.70	0.00	0.88	0.06
2	0.97	0.03	0.59	0.10	0.83	0.01	0.83	0.08
3	0.99	0.01	0.91	0.11	1.11	0.00	0.89	0.06
4	0.90	0.10	0.58	0.10	0.81	0.01	0.83	0.08
5	0.92	0.08	0.82	0.19	0.93	0.03	0.68	0.12
6	0.08	0.92	0.19	0.82	0.03	0.93	0.12	0.68
7	0.03	0.97	0.10	0.59	0.01	0.83	0.08	0.83
8	0.01	0.99	0.11	0.91	0.00	1.11	0.06	0.89
9	0.10	0.90	0.10	0.58	0.00	0.81	0.08	0.83
10	0.06	0.94	0.07	0.46	0.00	0.70	0.06	0.88
11	0.50	0.50	0.38	0.39	0.23	0.22	0.35	0.35
12	0.50	0.50	0.04	0.04	0.01	0.01	0.11	0.11
center 1	-2.99	0.54	-2.67	0.01	-3.03	0.02	-3.42	-0.00
center 2	2.99	0.54	2.67	0.01	3.03	0.02	3.42	-0.00

to that of PCM and PFCM as regards the two special points: point 11 is associated to values around 0.3, and point 12 to small values; it is thus identified as an outlier. Differences occur for other points: PCM and PFCM distributions are symmetric with respect to the cluster centre, whereas typicality is not. For instance, "extreme" points (points 1 and 10) have higher typicality degrees for their respective clusters (0.88) than "inner" points (e.g. points 5 and 6, associated to typicality 0.68). Indeed, the latter are considered as too similar to points of the other cluster and thus get lower external dissimilarity. This asymmetry property leads to a cluster repulsion effect: the cluster centres computed as weighted means of the data are further apart than centres obtained with other methods (see fig. 1 and the last two rows of table 2). This corresponds to prototypes that underline the cluster specificity and not only highlight the common features of the group members.

This property relates the typicality-based method to the algorithm proposed in [12], that adds to the PCM cost function a term modelling cluster repulsion, to solve the PCM cluster merging problem. This term then influences the centre expression. Typicality-based clustering also leads to this property, but in a different approach: cluster repulsion is not introduced in the centre expression, but directly in the point influence through its typicality degree.
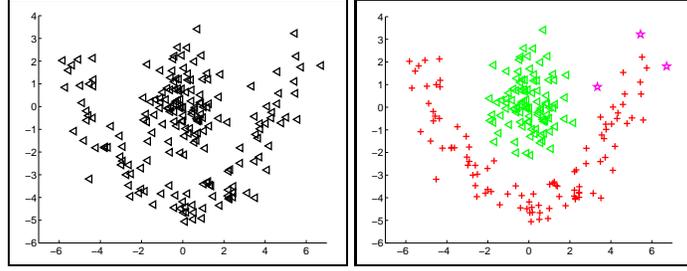
Figure 2. Considered data set and obtained non-convex clusters

5 Extension to Non-Convex Cluster

In this section, a direct extension of the proposed typicality-based clustering algorithm is presented to detect non-convex clusters: contrary to the fuzzy c-means and the possibilistic c-means algorithm for instance, the proposed method does not rely on the Euclidean distance. FCM and PCM optimise a cost function based on the Euclidean distance that leads to computing cluster centres as weighted means of the clusters members. Using other metrics requires specific adaptations [13].

On the contrary, the proposed typicality-based clustering does not compute a cluster centre, and does not assume the distance is Euclidean. It only relies on the definition of a resemblance and a dissimilarity measures, in a more general framework. Therefore, it can be applied to non-vectorial data using for instance kernel functions [5]. It can also be used to detect non-convex clusters, provided relevant comparison measures are defined.

Figure 2 illustrates this property for a data set made of a Gaussian cluster and a noisy parabolic cluster. We used a Gaussian kernel, i.e. the distance derived from the scalar product defined as

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

with $\sigma = 1.4$. It can be seen that the algorithm identifies the two expected clusters, considering as outliers (points assigned to the fictitious cluster, depicted with the star symbol) some points at the extremity of the parabola and a point located in an ambiguous area between the two clusters.

Figure 3 shows the associated typicality degree distribution as a function of the x-coordinate of the data points for each cluster respectively: they clearly indicate and distinguish between the two desired clusters, insofar as, for instance, points in the Gaussian cluster have a higher typicality degree for the first cluster than those in the parabolic cluster.
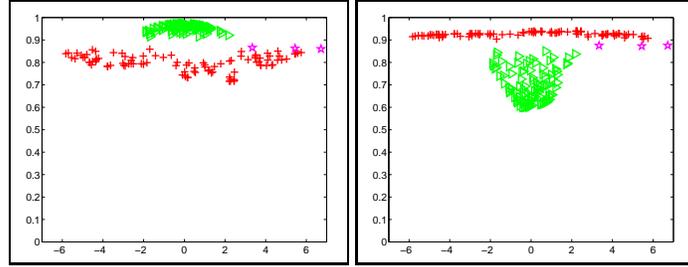
Figure 3. Typicality degrees for the Gaussian (left) and the parabolic (right) clusters, as a function of the x-coordinate of the data point (see the data set on fig. 2).

6 Conclusion

This paper presented a clustering algorithm based on typicality degrees originally defined in a prototype construction framework so as to take into account both common and discriminative features of categories to be characterised. Their exploitation for clustering makes it possible to identify relevant clusters that are both compact and separable, and provides information about the representativeness of the data points. It has the advantage of outlier robustness and cluster overlapping area avoidance. Furthermore, the algorithm is based on comparison measures and not on data points themselves, thus an extension to non Euclidean distances is possible, making it possible to identify non-convex clusters, or to apply the algorithm to non-vectorial data.

Perspectives include the classic problem of the cluster number selection in partitioning clustering methods. It would be interesting to study the extent to which classic validity criteria can be considered in this approach. Another interesting perspective concerns the extension to non-convex clusters and the definition of prototypes in this case: the computation of a weighted mean would not be relevant, as data are not compared in the initial input space, but in the so-called feature space to which data are implicitly transformed [11]. Thus cluster centres belong to the feature space and cannot be computed nor represented directly. The question of the definition of cluster representatives in the case of such kernel methods is an open question.

Acknowledgements

References

[1] Bouchon-Meunier B., Rifqi M., Bothorel S., 1996, *Towards general measures of comparison of objects*, Fuzzy sets and systems, Vol.84, No.2, pp. 143–153.

[2] Detyniecki M., 2000, *Mathematical aggregation operators and their application to video querying*. PhD thesis, Université de Paris VI.

[3] Krishnapuram R., Keller J., 1993, *A possibilistic approach to clustering*, IEEE Transactions on fuzzy systems, Vol.1, pp. 98–110.

[4] Lesot M.-J., 2005, *Similarity, typicality and fuzzy prototypes for numerical data*, 6th European Congress on Systems Science, Workshop "Similarity and resemblance".

[5] Lesot M.-J., Kruse R., 2006, *Data summarisation by typicality-based clustering for vectorial data and nonvectorial data*, Proc. of the IEEE Int. Conf. on Fuzzy Systems, Fuzz-IEEE'06, to appear.

[6] Lesot M.-J., Mouillet L., Bouchon-Meunier B., 2004, *Fuzzy prototypes based on typicality degrees* In Proc. of 8th Fuzzy Days, pp. 125–138.

[7] Pal N., Pal K., Keller J., Bezdek J., 2004, *A new hybrid c-means clustering model*, In Proc. of Fuzz-IEEE'04, pp. 179–184.

[8] Rifqi M., 1996, *Constructing prototypes from large databases*, In Proc. of IPMU'96.

[9] Rosch E., 1978, *Principles of categorization*, In: Rosch E. and Lloyd B. (Ed.), Cognition and categorization, Lawrence Erlbaum, pp. 27–48.

[10] Rosch E. and Mervis C., 1975, *Family resemblance: studies of the internal structure of categories*, Cognitive psychology, Vol.7, pp. 573–605.

[11] Schölkopf B., Smola A., 2002, *Learning with kernels*, MIT Press.

[12] Timm H., Borgelt C., Döring C., Kruse R., 2004, *An extension to possibilistic fuzzy cluster analysis*, Fuzzy Sets and Systems, Vol. 147, No.1, pp. 3-16.

[13] Zhang D., Chen S., 2003, *Clustering incomplete data using kernel-based fuzzy c-means algorithm*, Neural Processing Letters, Vol.18, No.3, pp. 155–162.