Title Page

Title	A Graphical method based on the Xie-Beni Validity index to improve the 'Possibilistic C-Means with Repulsion' Algorithm
Authors names	O.Shapira J.Wachs
Affiliation	Dept. of Industrial Engineering and Management, Ben-Gurion University of the Negev
Address	Dept. of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er Sheva 84105 Israel
phone	972-8-6472240
e-mail	{orensha, juan}@bgumail.bgu.ac.il

A Graphical method based on the Xie-Beni Validity index to improve the 'Possibilistic C-Means with Repulsion' Algorithm

Oren Shapira and Juan Wachs

Dept. of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er Sheva 84105 Israel

Abstract

The Possibilistic C-Means clustering algorithm, in its original form, is not very suitable for clustering due to the undesirable tendency to create coincident clusters and to converge to a "worthless" partitions in the case of poor initializations, but it provides robustness to noise and intuitive interpretation of the membership values. Recently, an extension of the PCM has been presented by Timm *et al.,* by introducing a repulsion term and showed a satisfactory performance when a proper value for the weighting factor γ was used. Our study has shown a correspondence between the Xie-Beni validity function and the range of the weighting factor of γ, and furthermore a practical graphical method and algorithm to find the suboptimal γ is presented. The results for the 'PCM with repulsion', compared to other possibilistic and probabilistic algorithms, showed quantitative superiority of the 'PCM with repulsion' over other methods.

Keywords: Possibilistic and probabilistic fuzzy clustering; Fuzzy C-Means; Cluster validity index, Robust methods.

1 Introduction

Cluster analysis is the process of classifying objects into subsets that have meaning in the context of a particular problem. The objects are thereby organized into an efficient representation that characterizes the population being sampled. (Jain and Dubes 1988). Hard clustering methods assume that each observation belongs to one class, however in practice clusters may overlap, and data vectors belong partially to several clusters. This scenario can be modeled properly using the fuzzy set theory (Zadeh 1965), in which the membership degree of a vector x_k to the i-th cluster (u_{ik}) is a value from [0,1] interval. Bezdek (1982) explicitly formulated this approach oriented to clustering by introducing the Fuzzy c-mean clustering algorithm. Unfortunately, this method showed the difficulty of high sensibility to noises and outliers in the data. To reduce this undesirable effect, a number of approaches have been proposed, but the most remarkable has been the possibilistic, introduced by Krishnapuram and Keller (1993), with their possibilistic c-means algorithm. In this algorithm the membership is interpreted as the compatibilities of the datum to the class prototypes (typicalities) which correspond to the intuitive concept of degree of belonging or compatibility. In the case of poor initializations, it is possible that the PCM will converge to a "worthless" partition where part or all the clusters are identical (coincident) while other clusters go undetected. Recently, a new scheme has been proposed, in order to overcome the problem of cluster mutual attraction forces, by introducing a supplementary term for cluster repulsion (Timm et al. 2001). By use of cluster repulsion, as good separation between clusters is obtained, as with the FCM, while keeping the intuitive concept and the noise insensibility introduced by the PCM. The goal of this paper is to establish a connection between the possibilistic approach and cluster validation indices, such that the quantitative superiority of the 'PCM with repulsion' over other methods is tangible.

The organization of this paper is as follows. In Section 2, we analyze the possibilistic approach by Krishnaparum and Keller (1993) developed to cope with the problem of noise and concept of compatibility, but lacks of clusters discrimination. In Section 3, we review the recently proposed method by Timm et. al. (2001) based on repulsion between clusters, and report the difficulty of choosing the proper value of the weighting factor γ. In Section 4, we compare four clustering techniques using several datasets and suggest a graphical method to obtain the optimal weighting factor γ; Finally, Section 5 presents our summary and conclusions.

2 Possibilistic Fuzzy Clustering

The most widely used prototype-based clustering method for data partition is probably the ISODATA or Fuzzy C-Means (FCM) algorithm (Bezdek 1982). Given a set of n data patterns, $X = x_1, ..., x_k, ..., x_n$., the algorithm minimizes a weighted within group sum of squared error objective function. A constraint assures relative numbers for the membership, and therefore is not suitable for applications where memberships are supposed to represent typicalities or compatibilities. Thus, in the FCM the memberships in a given cluster of two points that are equidistant from the prototype of the cluster can be significantly different and memberships of two points in a given cluster can be equal even thought the two points are arbitrarily far away from each other (Krishnapuram and Keller 1996). This situation is illustrated in Figure 1.

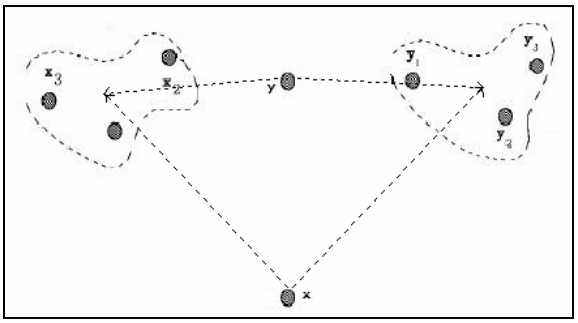

Figure 1. Example of dataset with a noise and outlier points

In this example, there are two clusters and a pair of points *x,y*; which represents an outlier and a noise point respectively. Intuitively, point y should not have a high degree value of membership (in the sense of compatibility) value for any cluster. Point *x*, should have an even smaller membership in either cluster, since it vaguely represents either one of them. Both points *x* and *y* will be assigned a membership value of 0.5 to both clusters by the FCM. One concludes that this membership values are unrepresentative of the degree of 'belonging', but also they cannot discriminate between an outlier datum and noise.

The PCM formulation relaxes the objective function of the FCM by dropping the sum to 1 constraint. To avoid a trivial solution of $u_{ik} = 0$ for all *i*, a penalty term is added which forces u_{ik} to be as large as possible, by modifying its objective function as follows:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} d^{2}(x_k, v_i) + \sum_{i=1}^{c} \eta_i \sum_{k=1}^{n} (1-u_{ik})^{m} \tag{1}$$

Where η_i is a positive number, and $u_{ik} \in [0,1]$. The new cluster centers v_i and the membership update equations are:

$$u_{ik} = \frac{1}{1+\left[\dfrac{d^{2}(x_k, v_i)}{\eta_i}\right]^{1/(m-1)}} \qquad v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} X_k}{\sum_{k=1}^{n} u_{ik}^{m}} \quad , \tag{2}$$

The parameter η_i is evaluated for each cluster separately; it determines the distance at which the membership degree equals 0.5.

$$\eta_i = K \sum_{k=1}^{n} u_{ik}^{m} d^{2}(x_k, v_i) \Big/ \sum_{k=1}^{n} u_{ik}^{m} \tag{3}$$

Using (3) η_i is proportional to the average fuzzy intracluster distance of cluster v_i. Usually K is chosen to be 1.

The main drawback of this promising approach appears when the objective function in (1) is truly minimized, and this occurs when all cluster centers are identical (coincident centroids). This failure is due to the reason that the membership degrees (2) depend only on the distance between the point to the cluster, and not on its relative distance to other clusters. Usually only part of the centroids are coincident, since the algorithm converge in a local minimum of the objective function (1), still this is an undesirable behavior for a clustering algorithm (Barni et. al. 1996).

3 Possibilistic Fuzzy Clustering with Repulsion

Recently, the Possibilistic Fuzzy Clustering with Repulsion was proposed to address the drawbacks associated with the FCM and the PCM. This method aims to minimize the intracluster distances (Bezdek and Pal, 1998), while maximizing the intercluster distances, without using implicitly the 'sum 1 restriction', but by adding a cluster repulsion

term to the objective function (1).

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} d^2(x_k, v_i) + \sum_{i=1}^{c} \eta_i \sum_{k=1}^{n} (1-u_{ik})^m + \gamma \sum_{i=1}^{c}\sum_{k=1,k\neq i}^{c} \frac{1}{d^2(v_i, v_k)} \quad (4)$$

Where γ is a weighting factor, and u_{ik} satisfies:

$$u_{ik} \in [0,1] \quad , \forall i \quad (5)$$

The repulsion term is relevant if the clusters are close enough. With growing distance it becomes smaller until it is compensated by the attraction of the clusters. On the other hand, if the clusters are sufficiently spread out, and the intercluster distance decreases (due to the first two terms), the attraction of the cluster can be compensated only by the repulsion term.

Minimization of (4) w.r.t. to cluster prototypes leads to:

$$v_i = \frac{\sum_{j=1}^{n} u_{ij} x_j - \gamma \sum_{k=1,k\neq i} v_k \frac{1}{d^2(v_k, v_i)}}{\sum_{j=1}^{n} u_{ij} - \gamma \sum_{k=1,k\neq i} \frac{1}{d^2(v_k, v_i)}} \quad (6)$$

Singularity occurs when one or more of the distances $d^2(v_k, v_i) = 0$ at any iterate. In such a case, (6) cannot be calculated. When this happens, assign zeros to each nonsingular class (all the classes except *i*) and assign 1 to class *i,* in the membership matrix *U.* Similar as for the PCM algorithm, the formula for updating the membership degrees u_{ik} is obtained using (2).

The weighting factor γ is used to balance the attraction and repulsion forces, i.e., minimizing the intradistances inside clusters and maximizing the interdistances between clusters. The central problem of this algorithm is that it requires a resolution parameter γ, and no clue is given about the correct range of this parameter (Dave and Krishnapuram, 1997).

Cluster validity studies the "goodness" of a partition generated by a clustering algorithm. The sum of intracluster distances, over the minimum of the intercluster distances is one of the most commonly used validity measures because of its analytical simplicity. This formulation is known as the Xie-Beni index v_{XB}, and is defined as:

$$v_{XB}(U,V;X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2 \| x_k - v_i \|^2}{n(\min_{i\neq j}\{\| v_i - v_j \|^2\})} \quad (7)$$

A good *(U,V)* pair should produce a small value of (7) because u_{ik} is expected to be high when $||x_k - v_i||$ and well-separated v_i's will produce a high value in the denominator of (7). Consequently, the minimum of v_{XB} , is the most desirable partition.

4 Tests Examples

The first example illustrates two well-separated noise-free clusters of 30 points each, drawn from two components, each one from a normal distribution, with μ_{1x}=2, μ_{1y}=3, μ_{2x}=5, μ_{2y}=3 and σ=0.5, see Figure 2.a. In Figure 2.b, the crisp partition for the FCM, PCM, FPCM, and 'PCM with repulsion' are similar, the centroids are almost the same. Each point is assigned to the cluster which it has the highest membership for the FCM, and for the possibilistic algorithms, the highest typicality was used. Ties are broken arbitrary. The parameters used were: *m*=2, *c*=2, ε=0.0001, η=2, γ=10.

After adding 28 points of random noise to one cluster (the cluster on the right of Figure 1.a) of the data set with μ_{2x}=6.5, μ_{2y}=4.5 and σ=1, the crisp partition presents significant differences; the FCM and FPCM performs the worst, while the other two present similar partitions, see Figures 2.c - 2.f. The FCM algorithm was used to obtain a good

initialization for the PCM. The noise addition to the original data affected significantly the cluster centers in the probabilistic based algorithms, while in the possibilistic based algorithms, the cluster centers are virtually unchanged.

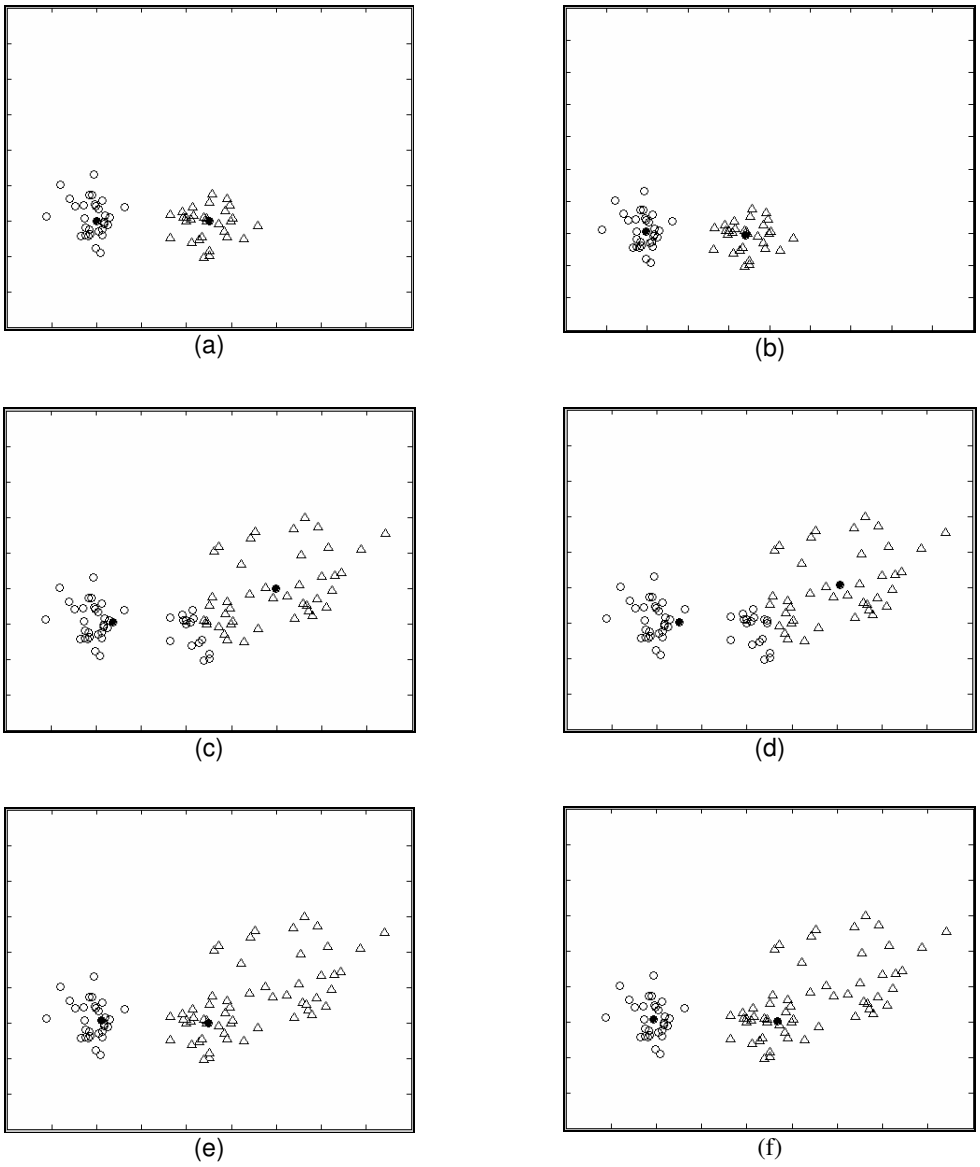

Figure 2. Partition of the synthetic data set: (a) the original data; (b) the FCM, FPCM, PCM and 'PCM with repulsion' partition; (c) the FCM partition with noise; (d) the FPCM partition with noise; (e) the PCM partition with noise; (f) the 'PCM with repulsion' partition with noisy data set.

Table 1, shows the centroids of the clusters, as seen by the different four algorithms. The last column presents the deviation of the centroids from their original location, after adding the noisy data. The performance of the PCM and 'PCM with repulsion' are acceptable. The FCM algorithm gave the poorest estimates for the centroids, since its deviation is the highest.

The second example shows a more realistic example with the well-known IRIS data set (Fisher 1936). IRIS consists of 150 points in four dimensions that represent three physical classes each with 50 points.
The numerical representation of two classes has substantial overlap, while the third is

well separated of the other two, therefore a three clusters classification is recommended. The algorithms discussed above have been tested on the IRIS data set, only both attributes, petal length and petal width have been used, since they carry the most information about the distribution of the iris flowers. Several runs of the four algorithms on IRIS data set were made, with different initializations and different (m, η, γ) tuples. Here we report results only for the initialization in Table 2. The performance of each algorithm (classification accuracy) is measured as the ratio between numbers of correct classification, and the total sample set. The number of mistakes is based on comparing the hardened version of the membership and typicalities matrices, to the physically correct crisp 3-partition of IRIS.

Table 1. Centroids estimation using the FCM, PCM, FPCM, and 'PCM with Repulsion' for Figure 2

	Fuzzy c-Means				Possibilistic				Mixed c-Means				Possiblistic with Repulsion			
Original Data Set	(1.89	3.08)	(4.47	2.91)	(1.96	3.04)	(4.41	2.93)	(1.89	3.08)	(4.47	2.91)	(2.01	3.06)	(4.37	2.93)
With Noise	(2.49	3.02)	(6.08	4.06)	(2.10	3.05)	(4.48	2.98)	(2.29	3.04)	(5.90	3.94)	(1.92	3.06)	(4.69	3.01)
Deviation	2.06				0.16				1.81				0.34			

Table 2. Classification accuracy on the IRIS data using the FCM, PCM, FPCM, and 'PCM with Repulsion'

Parameters			Accuracy				Iterations			
m	η	γ	FCM	FPCM	PCM	rep.PCM	FCM	FPCM	PCM	rep. PCM
2	3	0.1	82%	82%	64%	53%	26	12	26	13
		1				78%				
		15				97%				
		50				89%				
		100				76%				
		200				62%				

From Table 2, we find that the FPCM and 'PCM with repulsion' give higher or same accuracies than FCM for best γ cases, which indicate that typicality based classification represents better the physical data-partition, than membership values. Note that typicality-based classification accuracy failed for the PCM case, where two cluster coincidence affected the performance of the algorithm. The 'PCM with repulsion', using a good selection of parameter γ, shows significant superiority over the other algorithms studied here. Fortunately, Table 2, also shows that the number of iterations required for the PCM 'with repulsion' is similar to the FPCM, and is about half of that of the PCM and FCM.

Clusters detected by the PCM as a function of gamma, using m=2, are depicted in Figure 3. For γ=0.1 only two clusters are detected because the possibilistic algorithm is not forced to divide the data, and both clusters are coincident (Figure 3.a). By incrementing γ the attraction between clusters, decreases and the centroids of the coincident clusters are driven apart. For γ=1, the distance between samples assigned to clusters and their respective centroid is minimized, and that the distance between clusters is maximized, (Figure 3.b, Figure 3.c).

For further values of γ, the repulsion increases with the distance of the clusters, driving them ever farther apart, hence harming the classification accuracy (Figure 3.d). As shown in Table 2, in the 'rep.PCM' column, the parameter γ affect the performance of the data partition, and therefore, cluster validity measures should give some clue about the optimal value of γ.

For an unsupervised data set it is not possible to use the classification accuracy measure, and therefore an alternative method must be employed to find the best value of the weighting factor γ. We note that for the optimal value of γ, all the clusters are overall compact, and separate to each other.

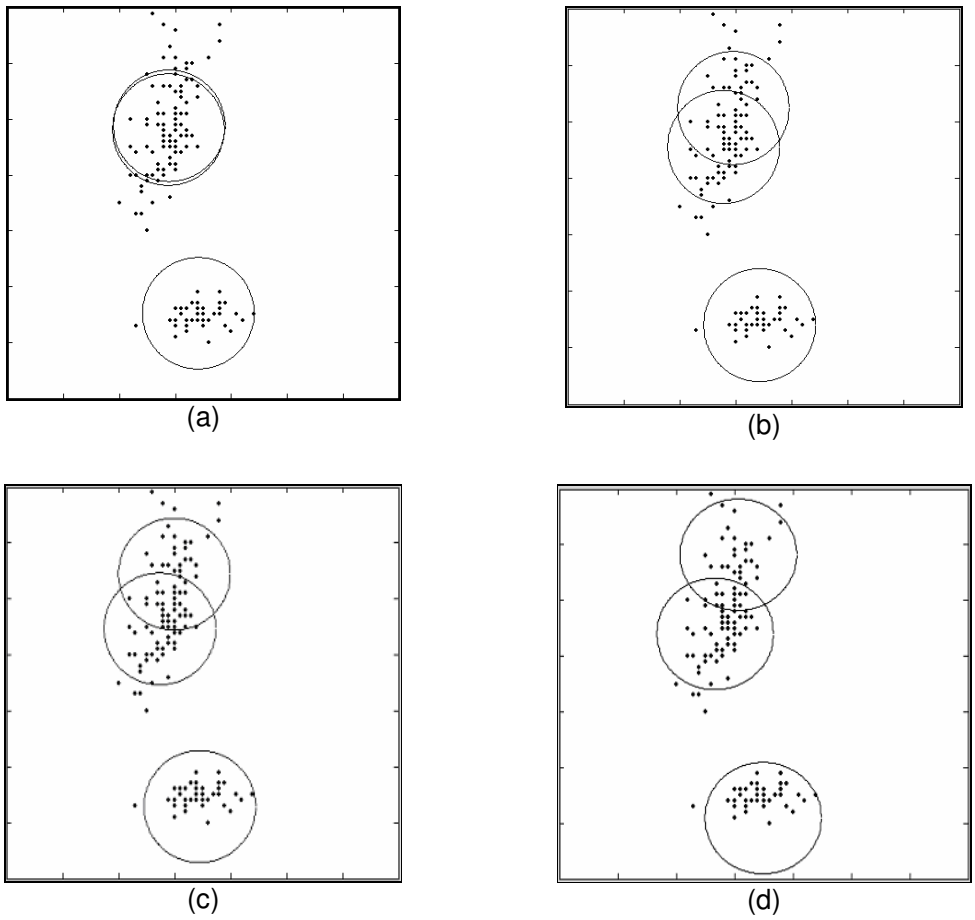

Figure 3. Iris dataset classified using 'PCM with repulsion': (a) γ=0.1, (b) γ=1, (c) γ=20, (d) γ=40.

A local minima in the compactness and separation validity function (7) indicates a better partition of the data set. For small values of γ, coincident centroids appear, then the denominator of (7) is negligible, and as a result of this, the v_{XB} is very high. However it, is noted that v_{XB} as a function of γ, is monotonically decreasing when γ gets very large (the denominator of (7) grows to infinity). Nevertheless, a close-up of the v_{XB} as a function shows that a local minimum is reached when the centroid is placed at its optimal value, see Figure 4.a. Still, if the centroid is misplaced far enough from all the others centroids, the points that should belong to clusters will be assigned to the closer centroid, and hence v_{XB} will drop abruptly. We conclude that the use of the Xie-Beni validation index is valid only around the optimal values for the centroids, where the validation function behaves convex.

The method proposed in this work is to find the optimal value of γ through plotting the function v_{XB} against γ, and graphically observing a "peak", see Figure 4. All the γ values immediately before and after the peak are candidates to be optimal. The γ* (optimal weighting factor) is the one which yields the lower v_{XB} between all the candidates. Figure 4.b, shows the plots for the hard partition of IRIS obtained using the 'PCM with repulsion', with m=2. The figure illustrates two valleys around the peak in γ=15, where the optimal between them is γ*=15.1, since there the Xie-Beni index is the lowest, v_{XB^*} =0.8. Observing Table 2, we see that a close to optimal classification accuracy (97%) was obtained for γ=15, therefore the result obtained graphically is satisfactory. This approach has been tested on the synthetic data set with and without noise, and we report results of 100% of classification accuracy using γ*=12 and γ*=15 respectively, see Figure 4.b-c. The last test has been conducted on the Wisconsin Breast Cancer Data (Merz and Murphy, 1996)

and a 95% of classification accuracy was obtained using γ*=14.9, see Figure 4.c. Once we have defined a method based on the validity function v_{XB}, our implementing strategy can be summarized into the following pseudo algorithm:

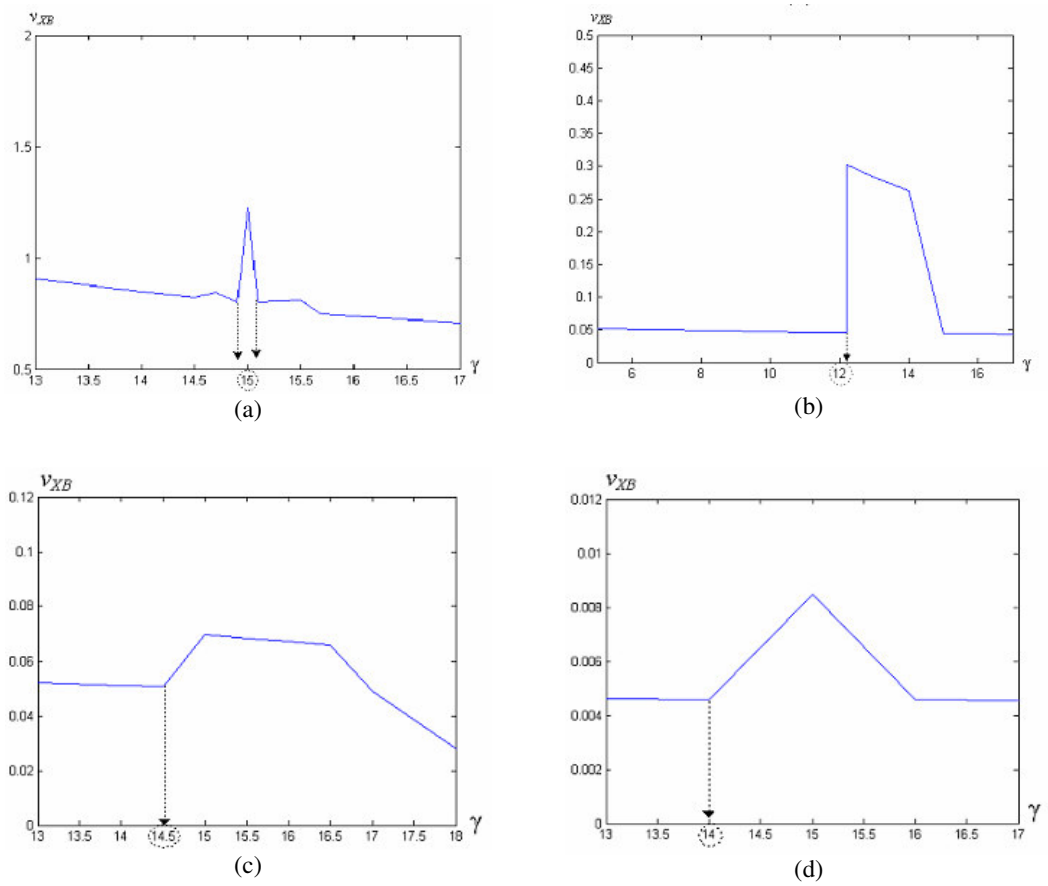

Figure 4. Plots of v_{XB} vs. γ : The monotonic decreasing function, a close-up peak for: (a) Iris data, (b) for synthetic data without noise (c) for synthetic data with noise and (d) for Wisconsin cancer breast data.

1) **Initialize** γ←0.2, Δ←0.1, m←2, K←1, $v_{XB}(U,V;X;\gamma-\Delta)\leftarrow 0$;
2) **Initialize** centroids v_i randomly;
3) Use (2) to **update** typicalities u_{ik} and to **update** centroids;
4) **Calculate** η_i using (3);
5) **Do** converge test; if negative goto 3;
6) **Compute** function $v_{XB}(U,V;X;\gamma)$;
7) **Repeat** steps 2,3,4,5 for $v_{XB}(U,V;X;\gamma+\Delta)$;
8) **If** $v_{XB}(U,V;X;\gamma-\Delta) < v_{XB}(U,V;X;\gamma) < v_{XB}(U,V;X;\gamma+\Delta)$ **then** $\gamma_k=\gamma$;
9) **If** γ=stop_value **then stop**; **else** γ= γ+2 Δ;
10) **Goto** 2;
11) **Find** $\gamma^*=min\ v_{XB}(U,V;X;\gamma_k)$ over all k;

Steps 2,3,4 and 5 are the 'PCM with repulsion' algorithm. The convergence test is $||J_{i+1} - J_i|| < \varepsilon$ where i is the number of iteration, J is the cost function obtained from (4) and ε is the accepted error margin. The fuzziness parameter m usually is set to 2, and the number of clusters c is fixed a priori. Initial values of v_i are datum selected randomly from the sample set.

The results for the 'PCM with repulsion', compared to other possibilistic and probabilistic algorithms, showed to be close to optimum when the weighting factor γ was obtained using the graphical method detailed in this paper. This method is suboptimal, however, computationally expensive in high dimensions. Its performance can be improved by considering the shapes of clusters and a better validity index.

5 Conclusions

We have proposed a graphical method to obtain a good range for the weighting factor γ, based on the Xie and Beni index validation measure, and proposed a strategy of implementation. Computational examples on two data sets were used to compare the four algorithms described in this paper, and to support the assertion that the weighting factor obtained by the graphical method presented is suboptimal. A limitation of the graphical method is that its visual identification of the sharp peaks is scale dependent and may be subjective. The points obtained using the validity index for discrete γ is a quantization of all the points on the curve of the validity function, and therefore the "peaks" might be missed. Nevertheless, when the quantization is fine, the accuracy of the estimates is rich. The tradeoff is that whenever the quantization of the weighting factor is finer, the search over the validity function is larger. Although the method proposed is computationally intensive, it is still better than supervising the clustering algorithm, especially when the labeling knowledge is unavailable.

Further investigation are required before much can be asserted about an optimal range for the weighting factor γ in presence of noisy environment which may cause false "peaks" in the validity function curve. Xie-Beni index only measures compact and separate clusters; therefore the reliability of the approach can be improved by considering better validity measures.

References

Bezdek, J. C., (1982), *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York.

Bezdek, J. C. and Pal, N. R. (1998), "Some New Indexes of Cluster Validity," *IEEE Trans. on System, Man, and Cybernetics, Part B*, vol. 28, no.3, pp. 301-315.

Barni, M., Cappellini V., and Mecocci, A. (1996), "Comments on 'A Possibilistic Approach to Clustering'," *IEEE Trans. Fuzzy Systems*, vol. 4, pp. 393-396.

Dave, R. N. and Krishnapuram, R. (1997), "Robust clustering methods: a unified view," *IEEE Trans. Fuzzy Systems,* vol. 5 no. 2, pp.270-293.

Fisher, R. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol.7, no. 2, 179-188.

Jain, A. K. and Dubes, R. C. (1988), *Algorithms for clustering data*, Prentice Hall, New Jersey.

Krishnapuram, R. and Keller, J. (1993), "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Systems, vol. 1, pp. 98-110.*

Krishnapuram, R. and Keller, J. (1996), "The Possibilistic C-Means algorithm: Insights and recommendations*". IEEE Trans. Fuzzy Systems*, vol. 4, pp. 385-393.

Merz, C. J. and Murphy, P. M. (1996), UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Department of Information and computer Science.

Timm, H., Borgelt, C. and Kruse, R. (2001), "Fuzzy cluster analysis with cluster repulsion," *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Tenerife, Spain.

Zadeh, L. A. (1965), "Fuzzy Sets," *Information and Control*, vol. 8, pp. 338-353.