Noise Clustering with a Fixed Fraction of Noise

Frank Klawonn

Department of Computer Science, University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany
f.klawonn@fh-wolfenbuettel.de

1 Introduction

Cluster analysis is an exploratory data analysis technique that is designed to group data and to detect structures within data. Exploratory techniques are applied as first steps in data analysis, immediately after elementary cleaning and visualisation of the data has been carried out. This means that a specific model for the data is not available at this state. However, exploratory methods usually incorporate control parameters that influence the result of this early data analysis step. Therefore, it is desirable to design methods that are robust w.r.t. the variation of such control parameters. In this paper, we modify the so-called noise clustering technique making it more robust against a wrong choice of its main control parameter, the noise distance. Section 2 briefly reviews the necessary background in fuzzy clustering. Section 3 introduces our modified noise clustering approach including a computationally efficient algorithm. We finish the paper by an example and some concluding remarks.

2 Objective Function-Based Fuzzy Clustering

Fuzzy clustering is suited for finding structures in data. A data set is divided into a set of clusters and – in contrast to hard clustering – a datum is not assigned to a unique cluster. In order to handle noisy and ambiguous data, membership degrees of the data to the clusters are computed. Most fuzzy clustering techniques are designed to optimise an object function with constraints. The most common approach is the so called probabilistic clustering with the objective function

$$f \;=\; \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij} \quad \text{constrained by} \quad \sum_{i=1}^{c} u_{ij} \;=\; 1 \quad \text{for all } j = 1,\ldots,n \quad (1)$$

that should be minimized. It is assumed that the number of clusters c is fixed. We will not discuss the issue of determining the number of clusters here and refer for an overview to [2, 4]. The set of data to be clustered is $\{x_1, \ldots, x_n\} \subset R^p$. u_{ij} is the membership degree of datum x_j to the ith cluster. d_{ij} is some distance measure specifying the distance between datum x_j and cluster i, for instance the (quadratic) Euclidean distance of x_j to the ith cluster centre. The parameter $m > 1$, called fuzzifier, controls how much clusters may overlap. The constraints lead to the name probabilistic clustering, since in this case the membership degree u_{ij} can also be interpreted as the probability that x_j belongs to cluster i. The parameters to be optimised are the membership degrees u_{ij} and the cluster parameters that are not given explicitly here. They are hidden in the distances d_{ij}. Since this is a non-linear optimisation problem, the most common approach to minimize the objective function (1) is to alternatingly optimise either the membership degrees or the cluster parameters while considering the other parameter set as fixed.

In this paper we are not interested in the great variety of cluster shapes (spheres, ellipsoids, lines, quadrics,...) that can be found by choosing suitable cluster parameters and an adequate distance function. (For an overview we refer again to [2, 4].) We only concentrate on the aspect of the membership degrees. Interpreting the membership degrees in terms of probabilities, u_{ij} specifies the probability that datum x_j belongs to cluster i, under the assumption that it must be assigned to a cluster. As a consequence, we obtain the following effect, which can lead to undesirable results. If we have, for instance, only two clusters and a datum has approximately the membership degree 0.5 to both clusters, it means either that the datum fits to both clusters equally well (the datum is near the border between the two clusters) or equally bad (the datum is noise and far away from both clusters).

In order to avoid this effect, possibilistic clustering was introduced [5], dropping the probabilistic constraint completely and introducing an additional term in the objective function to avoid the trivial solution $u_{ij} = 0$. However, the aim of possibilistic clustering is actually not to find the global optimum of the corresponding objective function, since this is obtained, when all clusters are identical. [6] describes an improved approach for the price of solving an additional non-linear optimisation problem in each iteration step.

Noise clustering [3] is another approach extending probabilistic clustering. The principle of probabilistic clustering is maintained, but an additional noise cluster is introduced. All data have a fixed (large) distance to the noise cluster. In this way, data that are near the border between two clusters still have a high membership degree to both clusters as in probabilistic clustering. But data that are far away from all clusters will be assigned to the noise cluster and have no longer a high membership degree to other clusters. The crucial point in noise clustering is the choice of the noise distance. If the noise distance is chosen too small, most of the data will simply be assigned to the noise cluster, if the noise distance is too high, the results are more or less identical to standard probabilistic clustering.

3 Noise Clustering with an Expected Fraction of Noise

The considerations in the previous section inspired the idea to introduce a new parameter into noise clustering that specifies the number or the fraction of noisy data expected or admitted in the data. At first sight, it seems that this approach makes the situation even more complicated, since in addition to the noise distance the additional parameter must also be specified. However, it turns out that a very rough (usually much too large) estimation of the noise distance in combination with a rough estimation of the number of noisy data leads to very good clustering results.

The basis of noise clustering with a fixed number $0 \leq N < n$ of noisy data is the following objective function

$$f = \sum_{i=1}^{c+1} \sum_{j=1}^{n} u_{ij}^2 d_{ij} \quad \text{constrained by} \quad \sum_{i=1}^{c+1} u_{ij} = 1 \quad \text{for all } j = 1, \dots, n. \tag{2}$$

$$\text{with the additional constraint} \qquad \sum_{j=1}^{n} u_{c+1,j} = N. \tag{3}$$

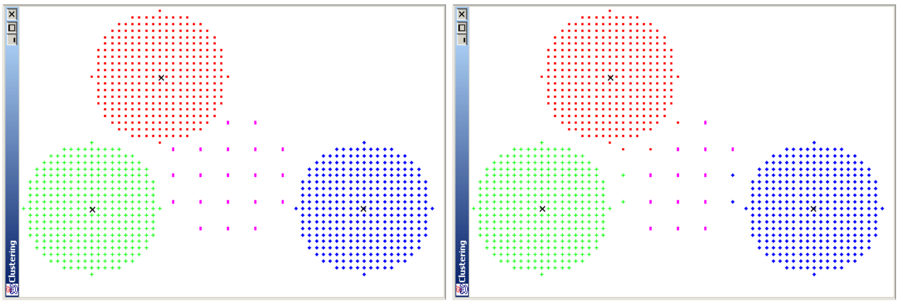

Fig. 1. Detection of noise with modified (left) and standard (right) noise clustering.

We assume that cluster number $(c+1)$ is the noise cluster and therefore the distance $d_{c+1,j} = d_{\text{noise}}$ is the fixed noise distance d_{noise}. In order to obtain an analytical solution for the alternating optimisation scheme, we have chosen a fuzzifier of $m = 2$. The constraint (3) reflects the requirement that N data are accepted or considered as noise. In order to obtain the update equations for the u_{ij}, we must find the global minimum of this objective function satisfying the constraints specified in (2) and (3), while we consider the distance values d_{ij} as fixed. Therefore, we compute the corresponding partial derivatives of the Lagrange function

$$f_L = \sum_{i=1}^{c+1} \sum_{j=1}^{n} u_{ij}^2 d_{ij} + \sum_{j=1}^{n} \lambda_j \left(1 - \sum_{i=1}^{c+1} u_{ij}\right) + \lambda \left(N - \sum_{j=1}^{n} u_{c+1,j}\right). \tag{4}$$

We obtain
$$\frac{\partial f_L}{\partial u_{ij}} \;=\; \left\{\begin{array}{ll} 2u_{ij}d_{ij} - \lambda_j & \text{if } 1 \le i \le c \\ 2u_{ij}d_{\text{noise}} - \lambda_j - \lambda & \text{if } i = c+1 \end{array}\right\} \;=\; 0. \qquad (5)$$

Writing down these equations and the constraints specified in (2) and (3), we obtain the following system of $(n(c+2)+1)$ linear equations.

$u_{1,1}$	$\dots$	$u_{1,n}$	$\dots$	$u_{c,1}$	$\dots$	$u_{c,n}$	$u_{c+1,1}$	$\dots$	$u_{c+1,n}$	λ_1	$\dots$	λ_n	λ	RHS
$2d_{1,1}$										-1				
	$\ddots$										$\ddots$			
		$2d_{1,n}$										-1		
			$\ddots$								$\vdots$			
				$2d_{c,1}$						-1				
					$\ddots$						$\ddots$			
						$2d_{c,n}$						-1		
							$2d_{\text{noise}}$			-1			-1	
								$\ddots$			$\ddots$		$\vdots$	
									$2d_{\text{noise}}$			-1	-1	
1				1			1							1
	$\ddots$				$\ddots$			$\ddots$						$\vdots$
		1				1			1					1
							1	$\dots$	1					N

Empty entries correspond to zeros. RHS stands for the right hand side of the equation. An ad hoc solution of this equation would not be feasible for large data sets. But we can see that the corresponding matrix is almost an upper triangular matrix. Only the last $(n+1)$ rows disturb the triangular structure. We can use the equations with $2d_{\text{noise}}$ to eliminate the coefficients 1 in the last row, thus replacing the last row by

$$\frac{1}{2d_{\text{noise}}}\lambda_1 + \ldots + \frac{1}{2d_{\text{noise}}}\lambda_n + \frac{n}{2d_{\text{noise}}}\lambda \;=\; N$$

or, equivalently,

$$\lambda_1 + \ldots + \lambda_n + n\lambda \;=\; 2Nd_{\text{noise}}. \qquad (6)$$

From (5), we know

$$u_{ij} \;=\; \frac{\lambda_j}{2d_{ij}} \quad (\text{for } i \le c) \qquad \text{and} \qquad u_{c+1,j} \;=\; \frac{\lambda_j + \lambda}{2d_{\text{noise}}}. \qquad (7)$$

The constraint in (2) together with (7) yields

$$\frac{\lambda_j + \lambda}{2d_{\text{noise}}} + \sum_{i=1}^{c} \frac{\lambda_j}{2d_{ij}} \;=\; 1, \qquad (8)$$

so that we have

$$\lambda_j = \frac{2 - \frac{\lambda}{d_{\text{noise}}}}{\frac{1}{d_{\text{noise}}} + \sum_{i=1}^{c} \frac{1}{d_{ij}}}. \tag{9}$$

Inserting (9) into (6) and solving for λ, we obtain

$$\lambda = \frac{2Nd_{\text{noise}} - 2\sum_{j=1}^{n} \frac{1}{\frac{1}{d_{\text{noise}}} + \sum_{i=1}^{c} \frac{1}{d_{ij}}}}{n - \frac{1}{d_{\text{noise}}} \sum_{j=1}^{n} \frac{1}{\frac{1}{d_{\text{noise}}} + \sum_{i=1}^{c} \frac{1}{d_{ij}}}}.$$

After having computed λ, we can use (9) to determine $\lambda_1, \ldots, \lambda_n$. Now we can directly compute the membership degrees u_{ij} from (7).

For this new type of noise clustering, the update equations for the membership degrees are no longer as simple as they are in probabilistic, possibilistic or standard noise clustering. However, the scheme we have derived is computationally efficient and is not significantly slower than the other clustering algorithms. If we had simply solved the system of linear equations in a naive way, the computation would not be feasible for larger data sets.

4 An Example

In this section we briefly illustrate with a simple data set, how our new algorithm works. Figure 1 shows the result of applying the well known fuzzy c-means clustering algorithm [1] with standard noise clustering on the right hand side and with our new approach on the left hand side. In both cases the cluster centres are positioned correctly and the data obviously belonging to a cluster are assigned with the highest membership degree to the corresponding cluster. The good result for standard noise clustering could only be obtained, by tuning the noise distance manually, finally to a value of 1.7. The data assigned to the noise cluster with the highest membership degree are marked as small vertically oriented rectangles. We can see that still some of the noisy data are not assigned to the noise cluster with the highest membership degree in the case of standard noise clustering. With our modification all noisy data are assigned to the cluster. For practical purposes, we do not specify the expected number of noisy data N, but the expected percentage P_{noise} or fraction of noisy data, so that the parameter N is determined by $N = \frac{P_{\text{noise}}}{100} \cdot n$. It should be noted that we can and should overestimate the percentage of noisy data. The nature of fuzzy clustering is that zero membership degrees nearly never occur. Therefore, even for a very large noise distance all data will have at least a small membership degree to the noise cluster. And all these small membership degrees contribute to the value of N of noisy data. Figure 2 (left) shows the clustering result with heavily overestimated noise where we have assumed 60% noisy data: $P_{\text{noise}} = 60$. Even for this case, the clustering result is still acceptable. In addition to the data that should be considered as noise, only data at the very boundary of the clusters are assigned to the noise cluster. If we apply standard noise clustering and decrease the more or less optimum noise distance of 1.7 in figure 1 to 1.6, the effect for the clustering is disastrous as the right hand side of figure 2 shows.

5 Conclusions

We have introduced an extension of noise clustering that allows the specification of the fraction of expected noisy data. A very rough value for this additional parameter frees the user from an accurate estimation of the noise distance. Therefore, our approach can be seen as a further step for making exploratory data analysis techniques more robust against tedious parameter selections.

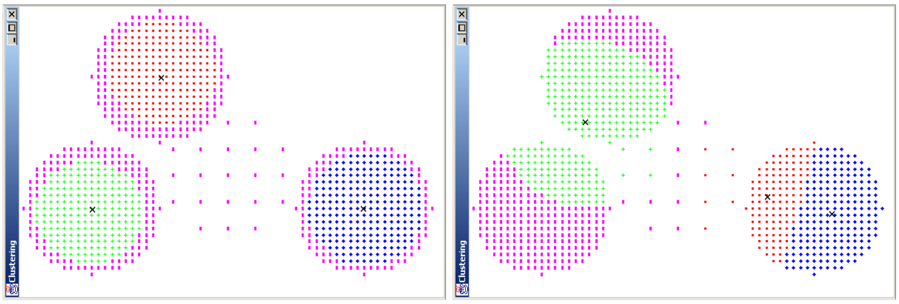

Fig. 2. Clustering result with heavily overestimated noise (left) and standard noise clustering with a slightly decreased noise distance.

References

1. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
2. Bezdek JC, Keller J, Krishnapuram R, Pal NR (1999) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer, Boston
3. Davé RN (1991) Characterization and detection of noise in clustering. Pattern Recognition Letters 12: 657–664
4. Höppner F, Klawonn F, Kruse R, Runkler T (1999) Fuzzy cluster analysis. Wiley, Chichester
5. Krishnapuram R, Keller J (1993) A possibilistic approach to clustering. IEEE Trans. on Fuzzy Systems 1: 98–110.
6. Timm H, Borgelt C, Kruse R (2002) A modification to improve possibilistic cluster analysis. IEEE Intern. Conf. on Fuzzy Systems, Honulul

Index