A Brief Overview of Robust Clustering Techniques

Olfa Nasraoui
Department of Computer Engineering & Computer Science
University of Louisville,
olfa.nasraoui_AT_louisville.edu

Robust Clustering

There are two major families of robust clustering methods. The first includes techniques which are directly based on robust statistics. Rousseeuw extended the idea of robust estimators to K clusters with his K-Medoid algorithm [RL87]. However this extension is based on the Median estimator (also known as L_1 regression) which minimizes the sum of absolute (instead of squared) residuals. For the case of estimating the parameters of a regression model, this estimator is resistant to outliers in the dependent variable, and enjoys a 50% breakdown point for location estimation. However, for regression with multivariate data, it is no better than regular LS in terms of its breakdown point (0%), because of its sensitivity to leverage points. The Generalized MVE (GMVE) estimator [JMB91] searches for clusters one at a time by repeatedly performing MVE followed by one-step of RLS after fixing the weights to 1 if a data point's Mahalanobis distance from the center is less than $\chi^2_{n,0.975}$ and 0 otherwise. In each pass, the Kolmogorov-Smirnov normality test is used to validate the fit of each cluster found before removing from the data set the points with weights equal to 1 in that cluster. Instead of fixing the fraction of inliers, h, to 50% as in the original MVE, GMVE searches all possible values of h within a prespecified interval with a suitable step size, to find each cluster. Hence, it is computationally quite expensive. GMVE has the advantage that the number of clusters need not be known in advance. However, it can not handle overlapping clusters.

The second family of robust clustering algorithms is based on modifying the objective of FCM to make the parameter estimates more resistant to noise. The first such attempt was presented by Ohashi [Oha84] who introduced the idea of a noise cluster indicated by the subscript (*), and modified the FCM criterion as follows

$$\min_{\mathbf{B},\mathbf{U}} \quad \alpha \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2} + (1-\alpha) \sum_{j=1}^{N} u_{*j}^{m}. \tag{1}$$

where α is a parameter related to scale. Later, in an independent effort, Davé proposed a Noise Clustering (NC) technique [Dav91] which uses a criterion similar to Ohashi's,

$$\min_{\mathbf{B},\mathbf{U}} \quad \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2} + \sum_{j=1}^{N} \delta^{2} \left(1 - \sum_{i=1}^{C} u_{ij}^{m}\right). \tag{2}$$

Both of these methods are easy to optimize. However, they had two major shortcomings. First, a single parameter is used to describe the scale or resolution parameter. This is clearly insufficient in the general clustering problem where clusters are not guaranteeed to be of the same size or to have the same inlier bounds. Second, the scale parameter needs to be known in advance, or pre-estimated from the data.

The Possibilistic C-Means (PCM) family of clustering algorithms was proposed by Krishnapuram and Keller to alleviate the noise problem by relaxing the constraint on memberships used in FCM [KK93,KK94]. It uses the objective function

$$\begin{aligned} J(\mathbf{B}, U, \mathcal{X}) \quad = \quad & \sum_{i=1}^{C}\sum_{j=1}^{N} (u_{ij})^{m} d^{2}(\mathbf{x}_{j}, \beta_{i}) + \\ & \sum_{i=1}^{C} \eta_{i} \sum_{j=1}^{N} (1 - u_{ij})^{m} \end{aligned} \tag{3}$$

where η_i are suitable resolution parameters which are allowed to vary from cluster to cluster, hence allowing clusters to have different sizes or inlier bounds. While the parameter update equations remain identical to those of FCM, it is easy to show [KK93] that U may be a global minimum of $J(\mathbf{B}, U, \mathcal{X})$ only if the memberships are updated by

$$u_{ij} = \frac{1}{1 + \left(\frac{d^{2}(\mathbf{x}_{j}, \beta_{i})}{\eta_{i}}\right)^{\frac{1}{m-1}}}. \tag{4}$$

Unlike the constrained FCM memberships, the Possibilistic memberships can be viewed as typicality values which measure how typical a point is of a given cluster, regardless of all other clusters.

It has been shown [NK95] that the PCM can be considered as a robust estimator representing C independent M-, and W-estimators [Hub81,RL87], with η_i being the scale parameter related to the spread of cluster β_i, and m being a parameter that determines the shape of the weight function. Since PCM tries to find the C best clusters independently of each other, it is possible that C identical clusters minimize the PCM objective function, while the remaining clusters are missed. Another problem that has proved to be a serious bottleneck for the PCM is that its performance relies heavily on a good initialization of the cluster parameters, and accurate estimates of η_i. When the initial prototype parameters or the resolution parameter η_i are inaccurate, the PCM can converge to a meaningless solution. In [KK93], it was suggested that the FCM algorithm be used to obtain the initial estimates of the prototypes, as well as the η_i. Several methods to estimate the η_i from the initial partition were proposed [KK93]. However, these methods can fail in the presence of noise, because if the data is noisy, the initial partition from the FCM may be too poor to yield good estimates of the η_i. Nasraoui and Krishnapuram used robust scale estimates along the lines of those used in robust statistics to initilialize the PCM in [NK96]. However, robust scale estimates still do not solve the problem of the dependence of these estimates on the accuracy of the initial prototype parameters. A better strategy to re-estimate the η_i in each iteration was also proposed in [NK96], albeit at a high computational cost, by using high breakdown scale estimates such as the median and the median of absolute deviations (MAD) [Hub81,RL87]. Also, it was recommended [NK95] that the membership function be modified so that it becomes redescending, i.e., it drops to zero beyond a finite rejection point.

Beni and Liu [BL94] proposed a Least Biased Fuzzy Clustering technique which minimizes for each cluster the clustering entropy defined by $\sum_{j=1}^{N} u_{ij} \log u_{ij}$, subject to the assumption that the centers are unbiased, i.e., $\sum_{j=1}^{N} u_{ij} \left(\mathbf{x}_j - \mathbf{c}_i\right) = \mathbf{0}$. The optimal centers are identical to those of the FCM, while the memberships are given by

$$u_{ij} = \frac{\exp\left\{-\beta d_{ij}\right\}}{\sum_{j=1}^{N} \exp\left\{-\beta d_{ij}\right\}}. \tag{5}$$

where β is a scale or resolution parameter resulting from the Lagrange multipliers used in the minimization process, and d_{ij} is the L_1 norm. This approach, like NC, suffers from the fact that a single scale parameter is used for all clusters, and this parameter has to be prespecified. It is also limited by the distance measure which is only suitable for spheroidal clusters.

Recently, Frigui and Krishnapuram [FK95] proposed the Robust $C-$ Prototypes (RCP) which explicitly incorporated M- and W-estimators into the FCM objective function as follows

$$\min_{\mathbf{B},\mathbf{U}} \quad \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^{m} \rho\left(d_{ij}^{2}\right). \tag{6}$$

subject to the same constraints on the memberships as in the FCM. The optimal prototype parameters β_i of the ith cluster are derived by setting $\frac{\partial J}{\partial \beta_i} = \mathbf{0}$ resulting in a system of p equations for each cluster similar to the W-estimator in ([*]), except that the contribution of each data sample inside the sum is multiplied by u_{ij}^{m}. RCP offers several innovative concepts such as the use of both a set of constrained fuzzy memberships to describe the degree of belonging of a data point to different clusters and a set of unconstrained robust weights for each cluster, that describe the adherence of a data point to the fit of that cluster as in W-estimators. However, RCP uses the Median and MAD to estimate the scale or resolution parameter used in the definition of the weights of each cluster. This requires a heavy computational cost, in addition to assuming that the noise proportion is at the most 50% in each cluster in each iteration.

Yager and Filev [YF94] proposed the mountain method which considers each data point as a candidate cluster center and chooses as an optimal cluster center, the data point at which a density-like function, $P(\mathbf{x}_i) = \sum_{j=1}^{N} \exp -\alpha \|\mathbf{x}_j - \mathbf{x}_i\|^2$, is maximized, where α is a resolution parameter. The algorithm proceeds by identifying one cluster at a time, and removing identified clusters by discounting their contribution to function $P(\mathbf{x}_i)$ of all the remaining points. The mountain method suffers from a very high polynomial complexity, $\mathcal{O}(N^2 + CN)$. It is also limited by its use of a single scale parameter, α, for all clusters, that is assumed to be known in advance. Davé and Krishnapuram [Da97] show the relation between many of the above-mentionned clustering algorithms, and discuss the role of validity in robust clustering.

Kim et al. [KKD95] proposed the Fuzy C Trimmed Prototypes algorithm (FCTP) which is based on the reformulated objective function of the FCM. The reformulated objective function of FCM is

$$\min_{\mathbf{B},\mathbf{U}} \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2 = \sum_{j=1}^{N} h_{ij}^2, \tag{7}$$

where $h_{ij}^2 = C\left[\sum_{i=1}^{C} \left(d_{ij}^2\right)^p\right]^{1/p}$, and $p = \frac{1}{1-m}$. The FTCP objective function is

$$J = \sum_{j=1}^{P} \left(h_i^2\right)_{j:N},$$

where $P < N$ and $(h_i^2)_{1:N} \leq \cdots \leq (h_i^2)_{N:N}$. They propose a heuristic algorithm to minimize J and estimate P. However, this procedure assumes a good initialization.

Choi and Krishnapuram [CK96] proposed a robust version of FCM based on the reformulated objective function of FCM. The objective function of this algorithm is given by

$$J = \sum_{j=1}^{N} \rho\left(h_j^2\right)$$

This algorithm uses only one robust weight per point as opposed to C weights in the formulation of (6). The weight is interpreted as the goodness of the point. This algorithm works only for spheroidal clusters. The choice of the loss function and the weights are not discussed in this paper.

We conclude from our review of existing robust estimation and clustering methods that they all suffer from the fact that they either depend on prespecified values for the scale parameters or the fraction of inliers, which makes them very sensitive to initialization; or they have to perform a quasi-exhaustive search on these parameters, which makes them require a very high computational cost.

Acknowledgment

This work is supported by the National Science Foundation (CAREER Award IIS-0133948 to O. Nasraoui). Partial support of earlier stages of this work by the Office of Naval Research grant (N000014-96-1-0439) and by the National Science Foundation Grant IIS 9800899 is also gratefully acknowledged.

Bibliography

BL94
G. Beni and X. Liu.
A least biased fuzzy clustering method.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(9):954-960, Sep. 1994.

CK96
Y. Choi and R. Krishnapuram.
Fuzzy and robust formulation of maximum-likelihood-based gaussian mixture decomposition.
In *IEEE Conference on Fuzzy Systems*, pages 1899-1905, New Orleans, Sep. 1996.

Da97
R. N. Davé and R. Krishnapuram and.
Robust clustering methods: A unified view.
IEEE Trans. Fuzzy Syst., 5(2):270-293, 1997.

Dav91
R. N. Davé.
Characterization and detection of noise in clustering.
Pattern Recognition Letters, 12(11):657-664, 1991.

FK95
H. Frigui and R. Krishnapuram.
A robust clustering algorithm based on the m-estimator.
In *Neural, Parallel and Scientific Computations*, Atlanta, Georgia, May 1995.

Hub81
P. J. Huber.
Robust Statistics.
John Wiley & Sons, New York, 1981.

JMB91

J. M. Jolion, P. Meer, and S. Bataouche.
Robust clustering with applications in computer vision.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(8):791-802, Aug. 1991.

KK93

R. Krishnapuram and J. M. Keller.
A possibilistic approach to clustering.
IEEE Trans. Fuzzy Syst., 1(2):98-110, May 1993.

KK94

R. Krishnapuram and J. M. Keller.
Fuzzy and possibilistic clustering methods for computer vision.
In S. Mitra, M. Gupta, and W. Kraske, editors, *Neural and Fuzzy Systems*, pages 135-159. SPIE Institute Series, 1994.

KKD95

J. Kim, R. Krishnapuram, and R. N. Davé.
On robustifying the c-means algorithms.
In *NAFIPS/ISUMA*, pages 630-635, College Park, MD, Sep. 1995.

NK95

O. Nasraoui and R. Krishnapuram.
Crisp interpretation of fuzzy and possibilistic clustering algorithms.
In *3rd European Congress on Intelligent Techniques and Soft Computing*, volume 3, pages 1312-1318, Syracuse NY, Aug. 1995.

NK96

O. Nasraoui and R. Krishnapuram.
An improved possibilistic c-means algorithm with finite rejection and robust scale estimation.
In *North American Fuzzy Information Processing Society Conference*, Berkeley, California, June. 1996.

Oha84

Y. Ohashi.
Fuzzy clustering and robust estimation.
Presentation at the 9th meeting of SAS Users Group International, 1984.

RL87

P. J. Rousseeuw and A. M. Leroy.
Robust Regression and Outlier Detection.
John Wiley & Sons, New York, 1987.

YF94

R. R. Yager and D. P. Filev.
Approximate clustering via the mountain method.
SMC, 24(8):1279-1284, 1994.

About this document ...

A Brief Overview of Robust Clustering Techniques

This document was generated using the **LaTeX2**HTML translator Version 2002 (1.62)

The command line arguments were:
latex2html `-split 0 -image_type gif RobustClustering.tex`