A Novel Approach to Noise Clustering for Outlier Detection

Frank Rehm[1], **Frank Klawonn**[2], **Rudolf Kruse**[3]

[1] German Aerospace Center, Braunschweig, Germany
e-mail: frank.rehm@dlr.de
[2] University of Applied Sciences Braunschweig/Wolfenbuettel, Germany
e-mail: f.klawonn@fh-wolfenbuettel.de
[3] Otto-von-Guericke-University of Magdeburg, Germany
e-mail: kruse@iws.cs.uni-magdeburg.de

Abstract Noise clustering, as a robust clustering method, performs partitioning of data sets reducing errors caused by outliers. Noise clustering defines outliers in terms of a certain distance, which is called noise distance. The probability or membership degree of data points belonging to the noise cluster increases with their distance to regular clusters. The main purpose of noise clustering is to reduce the influence of outliers on the regular clusters. The emphasis is not put on exactly identifying outliers. However, in many applications outliers contain important information and their correct identification is crucial. In this paper we present a method to estimate the noise distance in noise clustering based on the preservation of the hypervolume of the feature space. Our examples will demonstrate the efficiency of this approach.

Key words noise clustering, outlier detection, fuzzy clustering

1 Introduction

For many applications in knowledge discovery in databases finding outliers, rare events, is of importance. Outliers are observations, which deviate significantly from the rest of the data, so that it seems they are generated by another process [9]. Such outlier objects often contain information about an untypical behaviour of the system. However, outliers tend to bias statistical estimators and the results of many data mining methods like the mean value, the standard deviation or the positions of the prototypes of *k-means* clustering [6]. Therefore, before further analysis or processing of data is carried out with more sophisticated data mining techniques, identifying outliers is an important step.

Noise clustering (NC) is a method, which can be adapted to any prototype-based clustering algorithm like *k-means* and *fuzzy c-means* (FCM). The main concept of the NC algorithm is the introduction of a single noise cluster that will hopefully contain all noise data points. Data points whose distances to all clusters exceed a certain threshold are considered as noise. This distance is called the *noise distance*. These noisy points only have a relatively small effect on the mean calculation, which is, however, part of prototype-based clustering techniques. The crucial point is the specification of the noise distance δ.

In this work we present an approach to determine the noise distance based on the preservation of the hypervolume of the feature space when approximating the feature space by means of a specified number of prototype vectors.

After reviewing related work in the following section we describe the fuzzy c-means clustering algorithm in section 3, which is the basic clustering algorithm for the noise clustering technique which we will describe in section 4. In section 5 we finally present our approach for the estimation of the noise distance. By means of a simple example we will demonstrate the efficiency of our approach in section 6. Section 7 concludes with a brief outline on future work.

2 Related Work

Noise clustering has been introduced by Dave [2] to overcome the major deficiency of the FCM algorithm, its noise sensitivity. The noise clustering version of FCM will be explained in detail in section 3 and 4.

Possibilistic clustering (PCM) [12] is a method that controls the extension of each cluster by an individual parameter η_i. By means of η_i, PCM is widely robust to noise. However, it suffers from inconsistencies in the sense that instead of the global minimum of the underlying objective function a suitable non-optimal local minimum provides the desired clustering result. Such cases may occur because PCM-clusters sometimes overlap completely, since the algorithm does not imply any

control entity that prevents identical clusters. Similarities between PCM and noise clustering are analysed in [3], including a review of other methods regarding robust clustering.

In contrast to PCM where an individual distance η_i for each cluster defines some kind of noise distance, NC so far, classifies noise only on the basis of a global noise distance using one noise cluster. A generalised version of NC is introduced in [4]. A certain noise distance δ_j is assigned to each feature vector. This combination of NC and PCM performs partitioning and mode seeking [13] at the same time.

An approach to clustering in interaction with robust estimators is proposed in [11]. Clustering is followed here by alternating outlier detection and prototype adaptation, where the estimators can be applied attributewise to each single cluster.

3 Fuzzy Clustering

Most fuzzy clustering algorithms aim at minimising an objective function that describes the sum of weighted distances d_{ij} between c prototype vectors v_i and n feature vectors x_j of the feature space $\Re^p$:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m d_{ij}. \tag{1}$$

With the fuzzyfier $m \in (1,\infty]$ one can determine how much the clusters overlap. While high values for m lead to overlapping clustering solutions, small values, m tending to 1, lead to rather crisp partitions. In order to avoid the trivial solution assigning no data to any cluster by setting all u_{ij} to zero and avoiding empty clusters, the following constraints are required:

$$u_{ij} \in [0,1] \qquad 1 \le i \le c,\ \ 1 \le j \le n \tag{2}$$

$$\sum_{i=1}^{c} u_{ij} = 1 \qquad 1 \le j \le n \tag{3}$$

$$0 < \sum_{j=1}^{n} u_{ij} < n \qquad 1 \le i \le c. \tag{4}$$

When the Euclidian norm

$$d_{ij} = d^2(\mathbf{v}_i, \mathbf{x}_j) = (\mathbf{x}_j - \mathbf{v}_i)^T(\mathbf{x}_j - \mathbf{v}_i)$$

is used as distance measure for distances between prototype vectors v_i and feature vectors x_j, the fuzzy clustering algorithm is called *fuzzy c-means*. Modifications of the fuzzy c-means algorithm by means of the distance measure, i.e. by using the Mahalanobis distance, allow the algorithm to adapt different cluster shapes [7]. The minimisation of the functional (1) represents a nonlinear optimisation problem that is usually solved by means of Lagrange multipliers, applying an alternating optimisation scheme [1]. This optimisation scheme considers alternatingly one of the parameter sets, either the membership degrees

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{1}{m-1}}} \tag{5}$$

or the prototype parameters

$$\mathbf{v}_i = \frac{\sum_{j=1}^{n}(u_{ij})^m \mathbf{x}_j}{\sum_{j=1}^{n}(u_{ij})^m} \tag{6}$$

as fixed, while the other parameter set is optimised according to equations (5) and (6), respectively, until the algorithm finally converges. Nevertheless, the alternating optimisation scheme can lead to a local optimum. Therefore, it is advisable to execute several runs of FCM to ascertain a reliable partition. With the Euclidian distance measure the fuzzy c-means algorithm finds approximately equally sized clusters, which will be of interest for our noise clustering technique.

4 Noise Clustering

Fuzzy clustering with the fuzzy c-means algorithm allows, based on the membership degrees u_{ij}, the estimation of the degree of the assignment of a feature vector x_j to a prototype vector v_i. Since the sum of all membership degrees of a feature vector equals one according to equation (3), even outliers can achieve high membership degrees. Small membership degrees occur always due to border regions between two or more clusters. The idea of noise clustering is based on the introduction of an additional cluster, that is supposed to contain all outliers [3]. Feature vectors that are about the noise distance δ or further away from any other prototype vector get high membership degrees to this noise cluster. Hence, the prototype for the noise cluster has no parameters. Let v_c be the noise prototype and x_j the feature vector. Then the noise prototype is such that the distance d_{cj}, distance of feature vector x_j from v_c is the fixed constant value

$$d_{cj} = \delta^2, \quad \forall j.$$

The remaining $c-1$ clusters are assumed to be the good clusters in the data set. The prototype vectors of these clusters are optimised in the same way as mentioned in equation (6). The membership degrees are also adapted as described in equation (5). As mentioned above, the distance to the virtual prototype is always δ. The only problem is the specification of δ. If δ is chosen too small, too many points will get classified as noise, while a large δ leads to small membership degrees to the noise cluster, which means that noise data are not identified and have also a strong influence on the prototypes of the regular clusters.

5 Estimation of the Noise Distance

The specification of the noise distance depends on several factors, i.e. maximum percentage of the data set to be classified as noise [10], distance measure, number of assumed clusters and the expansion of the feature space. The noise distance proposed in [2] is a simplified statistical average over the non-weighted distances of all feature vectors to all prototype vectors

$$\delta^2 = \lambda \frac{\sum_{i=1}^{c-1} \sum_{j=1}^{n} d_{ij}}{n(c-1)}$$

where λ is the value of the multiplier used to obtain δ from the average of distances. As mentioned above one can show that in this way δ suffers from the fact that with an increasing number of prototypes δ assumes relatively high values. As a consequence, the placing of the prototypes will be effected by the outliers, which we intended to avoid.

We propose in this work a noise distance which depends primarily on the number of prototypes used for the clustering process and the expansion of the feature space. Under the constraint of the preservation of the hypervolume of the feature space, we chose for δ a value which corresponds to the cluster radius of the hyperspherical cluster. The cluster radius, so δ, will be chosen such that the sum of the hypervolumes of the $c-1$ good clusters with approximately same size, equals the hypervolume of the feature space. A uniformly distributed feature space would not have any outliers in this case. Consequently, if there are regions of high density, some prototypes will be attracted to these regions. Feature vectors which are located a larger distance away from any other prototype vector get high membership degrees to the noise cluster.

So the first step is to estimate the hypervolume of the feature space. A simple solution for this is shown in Figure 1(a). By means of the data set's extreme feature vectors the area of the resulting rectangle, or more generally, in an n-dimensional feature space, the hypervolume V of the cuboid, can be easily computed. A closer approximation of the hypervolume V can be achieved by subdividing the feature space into smaller pieces and summing up the single volumes of the respective hyperboxes. Figure 1(b) and 1(c) show two naive partitions of the feature space. Note, that such a grid should subdivide the feature space into hyperboxes of approximately the same size, since the fuzzy c-means algorithm searches for approximately equally sized clusters.

Assuming that clusters in the data set have approximately the same size, the cluster radius and the noise distance respectively is approximately the radius r of the hypersphere with a hypervolume about $V/(c-1)$, when using $c-1$ regular prototypes for the clustering. Since our estimation of the hypervolume is based on a rectangular shape, the radius of a corresponding hypersphere would not cover all feature vectors in the hyperbox. Furthermore, huge clusters may be approximated by several prototypes. The feature vectors in border regions of those prototypes should not get high membership degrees to the noise cluster. By all means, different applications require variable definitions regarding outliers. Thus, the noise distance δ can be tuned by a parameter α. Finally we obtain

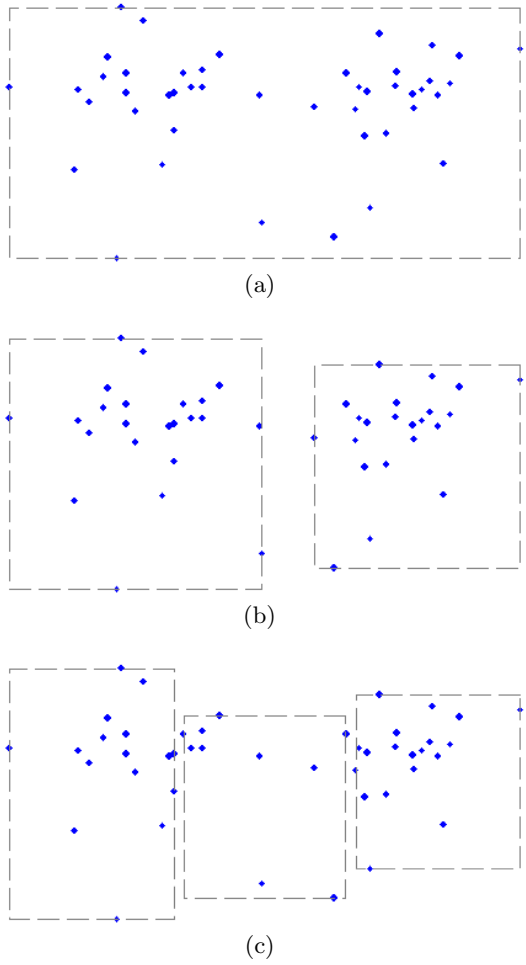


Fig. 1

$$\delta = \alpha r. \tag{7}$$

Although, any positive value can be chosen for α, our tests have shown that we achieve good results with $\alpha = 1.5$. In fact, smaller values of α lead to more compact clusters with a higher number of outliers. With $\alpha \to \infty$ NC tends to behave like FCM.

After defining the noise distance, we have to specify the minimum membership degree of a feature vector to the noise cluster in order to classify it as an outlier. It is obvious that no constant value will be appropriate to cover all NC partitions. Analogously to the noise distance, also the membership degree depends mainly on the number of prototypes used for the clustering. In [15] it is already discussed that the probability achieving high membership degrees with FCM, decreases with an increasing number of prototypes. The lower bound for

highest membership degrees is of course $1/c$. Of course, the noise distance affects the membership degrees to the regular clusters too, since a small noise distance forces high membership degrees to the noise cluster and small membership degrees to the regular clusters. So it makes sense to define outliers not only depending on the number of prototypes, but also by the fact what typical high membership degrees occur for a certain partition, which is naturally affected by the noise distance.

As we have discussed above, an outlier may be defined over the expected fraction of noise. With a simple method we can define outliers on the basis of the feature vector's probability not belonging to a regular cluster. Therefore, we estimate the mean value μ and the standard deviation σ of the membership degrees to the noise cluster and consider a feature vector as an outlier, if its membership degree to the noise cluster deviates more than a certain factor β from the mean value. Thus, the function $is\_outlier()$ returns 1 when, according to this definition, a feature vector is an outlier, otherwise the function returns zero:

$$is\_outlier(x_j) = \begin{cases} 1 & \text{if} \quad u_{cj} - \beta\sigma > \mu, \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

with

$$\mu = \frac{1}{n}\sum_{j=1}^{n} u_{cj} \tag{9}$$

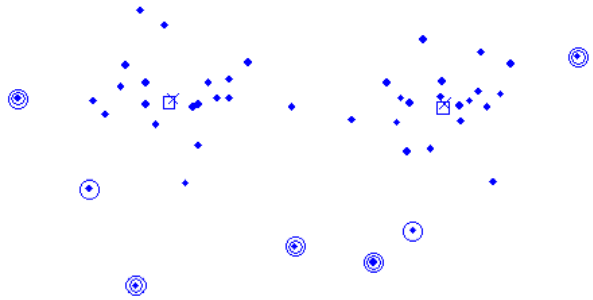
$$\sigma = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(u_{cj} - \mu)^2}. \tag{10}$$

Adjusting parameter β one can finally influence the fraction of outliers.

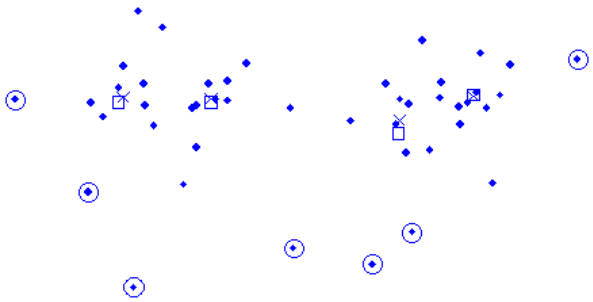
6 Experimental Results

Figure 2(a) shows the results of the two NC approaches. This data set obviously contains two clusters that are surrounded by some noise points (see also table 1). Using the conventional noise distance for partitioning the data set results in positioning the prototypes as plotted by squares in the figure. Applying the $is\_outlier()$ function with $\beta = 1.4$ the data points marked by the small circle are declared as outliers. The prototypes of the regular clusters are plotted in the figure with the $\square$ symbol. Since the conventional approach tends to overestimate the noise distance, the optimal cluster centres can not be found.

When we estimate the noise distance with our volume preserving approach, we obtain a much smaller value. Now the prototypes, that are plotted for this run with the $\times$ symbol, are placed closer to the respective cluster centre. In this way, two additional data points were identified as outliers, when we use the $is\_outlier()$ function again with $\beta = 1.4$. The outliers found with the new δ are marked by the bigger circle in the figure.



(a) Partitioning with 2 regular Prototpyes



(b) Partitioning with 4 regular Prototpyes

Fig. 2

Figure 2(b) shows the results on the same data set using four regular prototypes for the clustering. Real-life data sets usually contain cluster structures that differ from our assumption of hyperspheric clusters. The cluster structures must be approximated by several prototypes. A noise clustering technique should be able to deal with such challenges. As the figure shows, the conventional NC can not find any outlier in this example. This is the case, because the noise distance, when estimated by the conventional approach, will be approximately the same as for the two prototypes. But, as it can be easily verified by visual assessment, the distance from the prototypes to the representing data points is significantly smaller compared to partitioning with only two prototypes. Thus, when the average distance decreases with an increasing number of prototypes and the noise distance is almost constant, or even worse, increasing, then results similar to the one above will be obtained.

With our volume preserving approach, we obtain again the same result as with partitioning with two prototypes. The noise distance is, according to equation 7, much smaller. Therefore, the cluster centres can be placed better and distant points will be declared as outliers. The outliers found with the new δ are marked again by bigger circles in the figure.

x	y	x	y	x	y
-10.44	-1.33	36.04	-1.82	30.00	-18.00
14.75	-2.09	31.40	6.71	-20.00	-1.00
6.78	-1.03	38.79	5.04	15.00	-20.00
5.25	-0.87	26.78	1.07	25.00	-22.00
-3.84	1.09	29.72	-1.50	-5.00	-25.00
-1.29	8.42	33.74	1.28		
-6.86	0.60	28.51	-0.95		
-6.34	3.25	41.17	-0.40		
-4.47	10.40	42.47	3.50		
2.95	-1.70	36.18	-3.98		
9.21	3.65	27.98	-4.01		
6.74	1.47	38.29	-0.04		
-2.46	-4.25	22.22	-3.63		
-10.89	-12.67	32.33	-7.45		
1.19	-11.89	51.01	4.35		
-3.68	-1.73	37.20	-1.33		
-8.90	-3.05	29.25	-7.83		
2.24	-2.04	39.43	-1.97		
2.91	-7.08	33.58	-0.72		
4.28	1.14	40.18	-11.67		

Table 1 Synthetic data set (two clusters with some noise)

7 Conclusion

In this paper, we have proposed a new approach to determine the noise distance based on the preservation of the hypervolume of the feature space. Our approach is mostly independent of the number of clusters in the data set. Even though, we applied NC on FCM, other clustering algorithms, such as GK, GG and other prototype-based clustering algorithms [7,8] can be adapted. Our main concerns is not only to reduce the influence of outliers, but also to clearly identify them. Our results on the artificial example are promising. Subject of future work will be further methods to estimate the hypervolume of the feature space.

References

1. Bezdek JC (1980) A convergence theorem for the fuzzy ISODATA clustering algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 2: 1–8
2. Dave RN (1991) Characterization and detection of noise in clustering. Pattern Recognition Letters 12: 657–664
3. Dave RN, Krishnapuram R (1997) Robust clustering methods: a unified view. IEEE Transactions on Fuzzy Systems 5: 270–293
4. Dave RN, Sumit S (1997) On generalizing the noise clustering algorithms. Proceedings of the 7th Fuzzy Systems Association World Congress (IFSA-97) 3: 205–210
5. Dave RN, Sumit S (1998) Generalized noise clustering as a robust fuzzy c-m-estimators model. 17th Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS-98), Pensacola Beach, Florida, 256–260
6. Estivill-Castro V, Yang J (2004) Fast and robust general purpose clustering algorithms. Data Mining and Knowledge Discovery 8: 127–150, Kluwer Academic Publishers, Netherlands
7. Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 11: 773–781
8. Gustafson DE, Kessel WC (1979) Fuzzy clustering with a fuzzy covarianz matrix. Proceedings IEEE Conference on Decision and Control, San Diego, 761–766
9. Hawkins D (1980) Identification of outliers. Chapman and Hall, London
10. Klawonn F (2004) Noise clustering with a fixed fraction of noise. In: Lotfi A, Garibaldi JM (eds) Applications and Science in Soft Computing, Springer, Berlin, 133–138
11. Klawonn F, Rehm F (2006) Clustering techniques for outlier detection. In: Wang J (ed) Encyclopedia of Data Warehousing and Mining, Idea Group, Hershey, 180–183
12. Krishnapuram R, Keller JM (1993) A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1: 98–110
13. Krishnapuram R, Keller JM (1996) The possibilistic c-means algorithm: insights and recommendations. IEEE Transactions on Fuzzy Systems 4: 385–393
14. Santos-Pereira CM, Pires AM (2002) Detection of outliers in multivariate data: a method based on clustering and robust estimators. In: Härdle W, Rönz B (eds) Proceedings in Computational Statistics: 15th Symposium held in Berlin, Physica-Verlag, Heidelberg, 291–296
15. Windham MP: Cluster validity for fuzzy clustering algorithms. Fuzzy Sets and Systems 5: 177–185