Journal of Global Optimization (2006) 35: 625–635
DOI 10.1007/s10898-005-5372-5

Data Clustering Based on Maximization of Outlier Factor

VYDUNAS SALTENIS
Institute of Mathematics and Informatics, Akademijos 4, LT-08663 Vilnius, Lithuania (e-mail: saltenis@ktl.mii.lt)

(Received 17 November 2005; accepted 21 November 2005)

Abstract. There exist many data clustering algorithms, but they can not adequately handle the number of clusters or cluster shapes. Their performance mainly depends on a choice of algorithm parameters. Our approach to data clustering and algorithm does not require the parameter choice; it can be treated as a natural adaptation to the existing structure of distances between data points. The outlier factor introduced by the author specifies a degree of being an outlier for each data point. The outlier factor notion is based on the difference between the frequency distribution of interpoint distances in a given dataset and the corresponding distribution of uniformly distributed points. Then data clusters can be determined by maximizing the outlier factor function. The data points in dataset are divided into clusters according to the attractor regions of local optima. An experimental evaluation of the proposed algorithm shows that the proposed method can identify complex cluster shapes. Key advantages of the approach are: good clustering properties for datasets with comparatively large amount of noise (an additional data points), and an absence of important parameters which adequate choice determines the quality of results.

Key words: clustering, global optimization, local optimization, outlier detection

1. Introduction

Cluster analysis (Jain and Dubes, 1988) divides data into groups of similar objects. Each group consists of objects that are in a sense similar between themselves and dissimilar to objects of other groups. Clustering requires the definition of a similarity measure between patterns, which is not easy to specify in the absence of knowledge about cluster shapes.

A large number of clustering algorithms exist (Jain et al., 1999); each algorithm has its own approach for handling number of clusters, their shape, and structure of the data. Clustering techniques are divided in *hierarchical* and *partitioning*. Hierarchical algorithms build clusters gradually, and partitioning algorithms detect clusters directly trying to identify clusters as areas highly populated with data. Partitioning algorithms are less sensitive to outliers and can discover clusters of irregular shapes.

The K-means algorithm is one of the simplest clustering algorithms. Its limitation is inability to identify clusters with arbitrary shapes. K-means

methods are not very stable and very sensitive to outliers. They often do not work well when the clusters are of different size, shape, and density (Ertoz et al., 2002).

One of the problems with the clustering methods is that the most clustering algorithms prefer certain cluster shapes, and the algorithms will assign the data to clusters of such shapes even if there were no clusters in the data.

Another problem is that the choice of the number of clusters may be critical: different clusters may emerge when the number of clusters is changed. Good initialization of the cluster centroids in a K-means clustering method (MacQueen, 1967) may also be crucial; some clusters may even be left empty.

Hinneburg and Keim (1998) used density functions defined over the attribute space. They proposed the algorithm DENsity-based CLUstEring (DENCLUE) which is based on the idea that the influence of each data point can be modeled using some influence function. The overall density function of the data space can be calculated as the sum of the influence functions of all data points. This function is multimodal; each maximum corresponds to the cluster center. Clusters can be determined by identifying density attractors after local optimization of overall density function for each data point.

The influence function in Hinneburg and Keim (1998) can be an arbitrary function, for example:

- Square Wave Function

$$f_{\text{Square}}(x, y) = \begin{cases} 0, & \text{if } d(x, y) > \sigma, \\ 1, & \text{otherwise.} \end{cases}$$

- Gaussian Influence Function

$$f_{\text{Gauss}}(x, y) = \exp\left(-\frac{d(x, y)^2}{2\sigma^2}\right),$$

where $d(x, y)$ is the distance between two vectors.

The results of the algorithm mainly depend on a choice of the influence function and its parameters (for example, parameter σ). But the idea of data clustering by local optimization of some function defined over the attribute space is successfully used in this paper. The outlier factor introduced in Saltenis (2004) does not require the parameter choice and was used as some overall density function. The outlier factor can be seen as natural adaptation to the existing structure of distances between data points.

The rest of the paper is organized as follows. In section 2 outlier factor and outlier factor function are introduced. In section 3 the problems of optimization of outlier factor function are discussed. In section 4 the

proposed optimization procedure is presented. In section 5 we provide an experimental evaluation of our approach. Section 6 summarizes the results.

2. Outlier Factor and Outlier Detection

Outliers and clusters in a dataset are related. An outlier means not being in or close to a cluster. The outlier factor (Saltenis, 2004) used in this paper achieves minimal values for outlier points and maximal values in cluster centers. A useful idea to evaluate the outlier factor is to analyze the distribution of distances between the points.

The authors (Brin, 1995; Steinbach et al., 2003) paid an attention, that one way of analyzing whether a data may contain clusters is to plot the approximate probability density function of the pairwise distances between all points in a dataset. If the data contains clusters then the histogram will show two peaks: a peak representing the distance between points in clusters and a peak representing the average distance between the points (see Figure 1). If only one peak is present then clustering will likely be difficult.

One way to use these distributions with the aim to extract some properties of point outlierness is to compare them with the corresponding distributions of uniformly distributed points (see Figure 2).

The tables of the corresponding distribution values evaluated experimentally are presented in Saltenis (2004).

We can see in all cases the dominating narrow peak, which is usual for uniform distribution of data points in multidimensional cube. This peak is different for different space dimensionality.

The main idea is to eliminate the influence of dominating narrow peak and analyze the difference between the distribution of the pairwise distances d between all points in a given dataset $f^n(d)$ and the corresponding distribution of uniformly distributed points $f^u(d)$ for the same dimensionality. The analyzed domain is the same multidimensional cube in both cases.

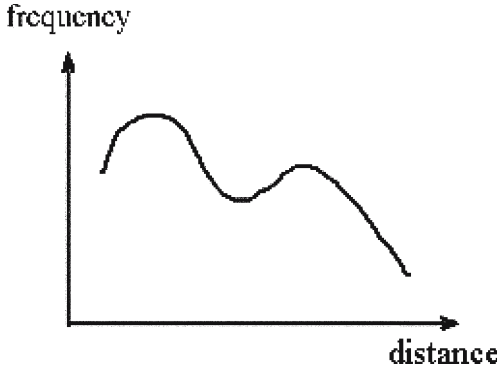

Figure 1. Plot of interpoint distances for data with clusters.

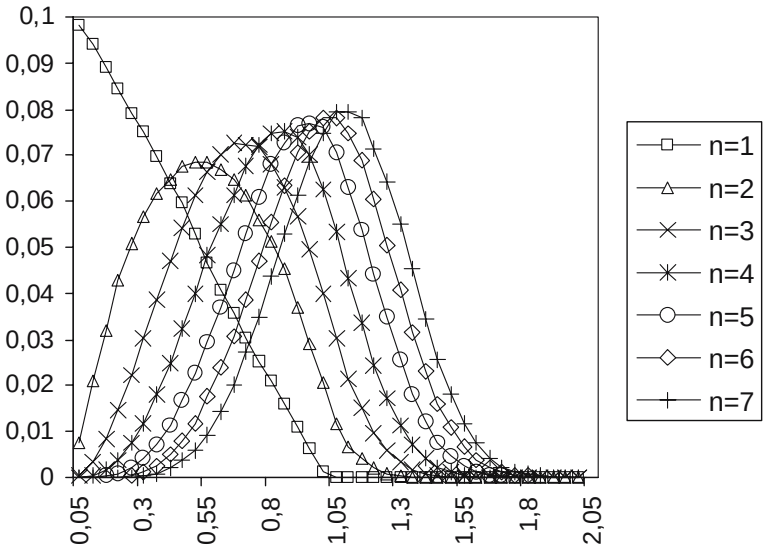

Figure 2. The frequency distributions of uniformly distributed points for dimensionalities $n = 1, \ldots, 7$ in a hypercube of side length 1.

The difference function

$$f(d) = f^n(d) - f^u(d)$$

may be treated as a *frequency function*, similar to the influence function introduced in Hinneburg and Keim (1998).

If points are uniformly distributed we obtain frequency function that is near to zero to all interpoint distances.

If the points are not uniformly distributed the greatest positive values of frequency function will indicate the most frequent, the most typical interpoint distances.

Figure 3 gives an illustration of frequency function for the data points in two-cluster situation (dimensionality 2) in Figure 4. The first peak of positive values of the frequency function is due to distances inside the point clusters, and the frequent distances between the clusters cause the second peak of positive values.

For each data point $i = 1, \ldots, m$ an *outlier factor* may be calculated:

$$R_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} f(d(X_i, X_j)),$$

where $X = (x_1, \ldots, x_n)$, and $d(X_i, X_j)$ is the distance between two vectors.

The outlying points will have low values of outlier factor R because the distances between the point and the rest points will be a typical.

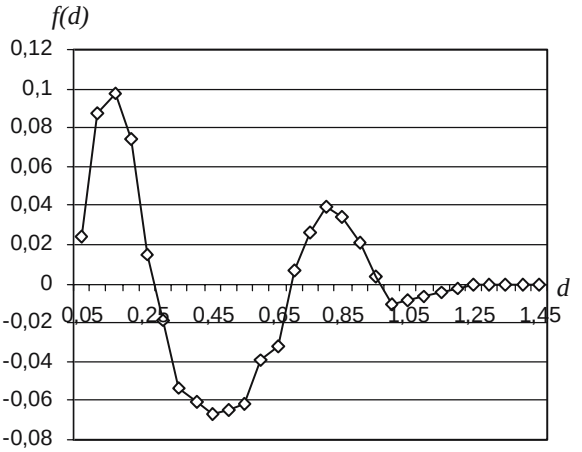

Figure 3. Frequency function for the data point allocation of Figure 4.

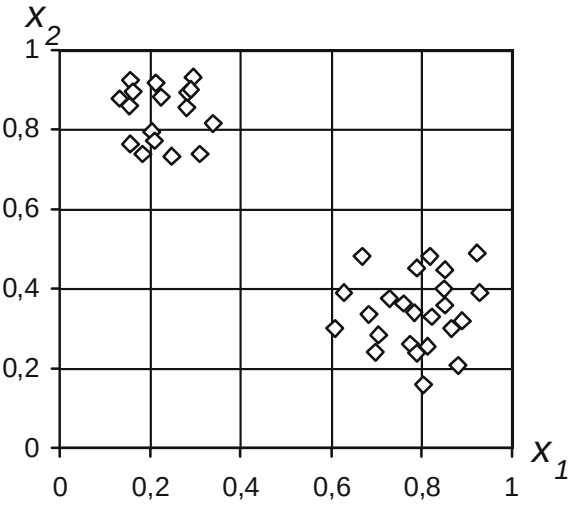

Figure 4. Data point allocation situation for two clusters.

The factor was used to rank the dataset objects regarding their degree of being an outlier (Saltenis, 2004). To investigate the quality of the outlier detection, the experiments were performed on widely used HBK (Hawkins et al., 1984) and Wood (Draper and Smith, 1966) datasets. A comparison with some popular detection methods demonstrated the superiority of the approach.

In the same way as outlier factor we also may introduce an *outlier factor function* $R(X)$ for each space point X:

$$R(X) = \sum_{j=1}^{m} f(d(X, X_j)).$$

The illustrations of outlier factor function values together with the corresponding data points for two-dimensional data are presented in Figures 5 (two clusters) and 6 (three clusters).

3. Optimization of Outlier Factor Function

It is obvious that if local optimization procedure of the outlier factor function $R(X)$ converges to the same local maxima when start points of the optimization are data points X_i and X_j then these data points belong to the same cluster. Then points of local maxima are the centers of clusters, and the point of global maximum is the center of dominant cluster.

Data clustering experiments using second order optimization method demonstrated some drawbacks of this approach. Optimization results may *slightly* differ for the points of the same cluster because the outlier factor function may have relatively small local optima near the data points, and the function is not always continuous or differentiable. Figure 7 presents an illustration of two local maxima of outlier factor function in case of one data cluster.

For this reason more stable heuristic optimization procedure was proposed.

4. Heuristic Optimization Procedure

The procedure uses only outlier factor values R_i of data points X_i, $i = 1, \ldots, m$.

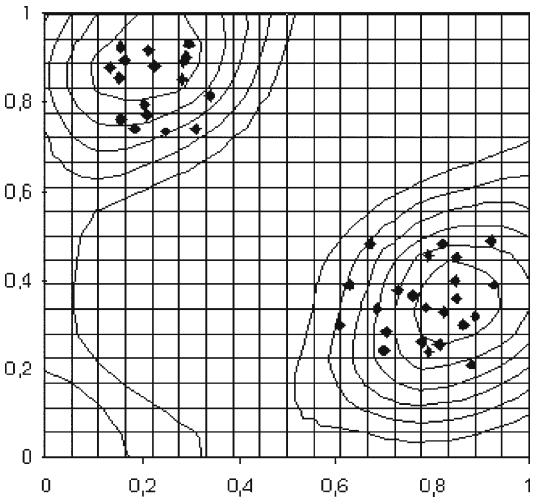

Figure 5. The outlier factor function together with the corresponding two-dimensional data points (two-cluster case).

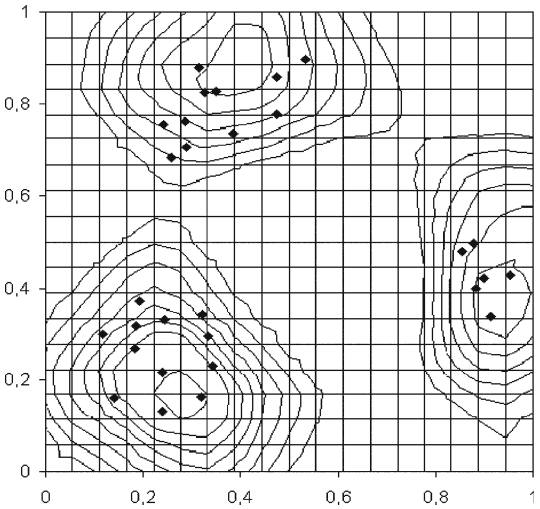

Figure 6. The outlier factor function together with the corresponding two-dimensional data points (three-cluster case).

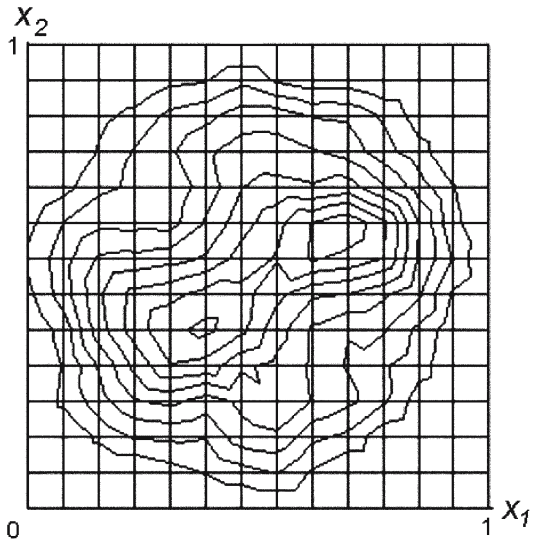

Figure 7. Illustration of two local maxima of outlier factor function.

1. Each data point is used as a start point for local optimization procedure. If optimization procedure converges to the same local maxima for different start points X_i and X_j then these start points belong to the same cluster.
2. In each step of optimization procedure current data point is changed to the nearest data point which outlier factor value is greater.

3. If outlier factor values for all points in some distance ε from the current point are not greater then this point is the local optimum point, cluster center.
4. The distance ε used for the limitation of search for better point is the only parameter of procedure.

The values of parameter ε may slightly influence the results of clustering. In our experiments the parameter value used was equal to 0,1 for hypercube domain of side length 1.

In the next section, we present the illustration of the optimization procedure steps for Iris data.

5. Experimental Results

We illustrate the experimental results of the proposed clustering method with several test datasets.

The half-rings dataset, as shown in Figure 8 consists of two clusters (20 points in each cluster).

Clustering was performed properly; two clusters were selected. Black marks in Figure 8 were used for cluster centers, and x marks were used for the outliers in each cluster. The K-means algorithm was unable to identify the two natural clusters, imposing a spherical structure on the data.

Iris data (Fisher, 1936) has well-known structure: the first cluster (50 data points) is well separated, and there are two contiguous clusters (50 points in each). Four attributes define these data.

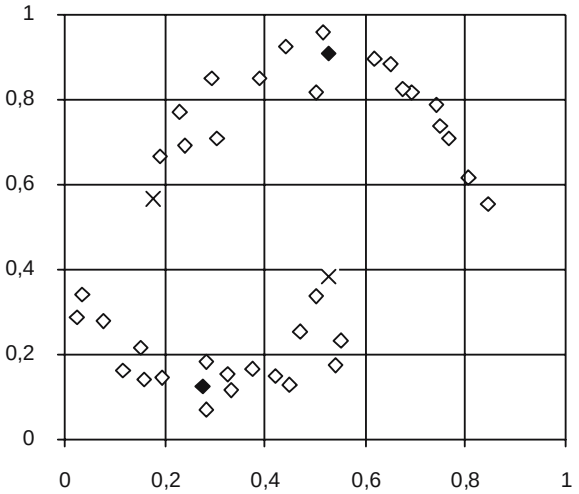

Figure 8. The half-rings dataset consisting of two clusters.

The data of first cluster were separated properly. The points of second and third clusters were also separated properly with some exceptions. One data point from the third cluster was attached to the second cluster, and 13 data points from second cluster were attached to the third cluster.

Figure 9 presents an illustrative scheme of heuristic optimization procedure steps for data points of second cluster.

We can see the numbers of data points and the sequence of the points in optimization steps. The end of all optimization procedures with various start points is data point no 74. This point is the center of the second cluster of Iris data set.

Noise invariance was investigated in such a way:

- Two gaussian data clusters without noise were used.
- The distance between them was selected so that the proposed technique could identify the clusters.
- Additional uniformly distributed data points (the noise) step by step were introduced and each time cluster analysis was performed.

Results of the investigation are presented in Table 1.

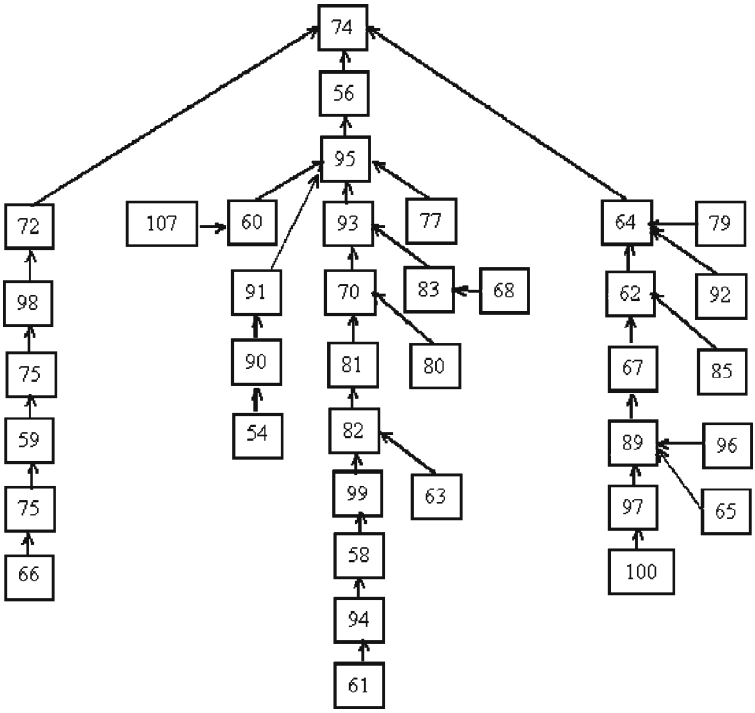

Figure 9. Illustrative scheme of heuristic optimization procedure steps for second cluster of Iris data points.

Table I. Results of noise invariance investigation

Percent of additional noisy data points	Quality of the clustering
0–55%	No errors, two clusters
55–62%	55% of errors, two clusters
> 62	One cluster

6. Conclusions

The outlier factor and outlier factor function may be successfully used in data clustering.

The advantages of new approach and optimization procedure are:

- good clustering properties for data sets with large amount of noise;
- the absence of important parameters which choice determines the quality of results;
- evaluation of the frequency function can be seen as natural adaptation to the dataset.

Acknowledgments

The research was partially supported by the Lithuanian State Science and Studies Foundation, Grant No. C 03013.

References

1. Brin, S. (1995), Near Neighbor Search in Large Metric Spaces. In: *Proceedings of the 21st International Conference on Very Large Databases (VLDB-1995)*, Zurich, Switzerland, Morgan Kaufmann, pp. 574–584.
2. Draper, N.R. and Smith, H. (1966), *Applied Regression Analysis*, Wiley, New York.
3. Ertoz, L., Steinbach, M. and Kumar, V. (2002), A new shared nearest neighbor clustering algorithm and its applications, AHPCRC, Technical Report 134.
4. Fisher R.A. (1936), The use of multiple measurements in taxonomy problems, *Annals of Eugenics* 7, 179–188.
5. Hawkins, D.M., Bradu, D. and Kass, G.V. (1984), Location of several outliers in multiple regression data using elemental sets, *Technometrics* 26, 197–208.
6. Hinneburg, A. and Keim, D. (1998), An efficient approach to clustering large multimedia databases with noise. In: *Proceedings of the 4th ACM SIGKDD*, New York, NY, pp. 58–65.
7. Jain, A.K. and Dubes, R.C. (1988), *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ.
8. Jain, A., Murty, M.N. and Flynn, P. (1999), Data clustering: a review. *ACM Computing Surveys* 31(3), 264–323.
9. MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M. and Neyman, J. (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume I: Statistics*, University of California Press, Berkeley and Los Angeles, CA, pp. 281–297.

10. Saltenis, V. (2004), Outlier detection based on the distribution of distances between data points, *Informatica*, 15(3), 399–410.
11. Steinbach, M., Ertoz, L. and Kumar, V. (2003), *Challenges of Clustering High Dimensional Data. New Vistas in Statistical Physics. Applications in Econophysics, Bioinformatics, and Pattern Recognition*, Springer-Verlag, Berlin.