Data clustering for circle detection

Tomislav Marošević
Data clustering for circle detection

Tomislav Marošević*
Department of Mathematics
J. J. Strossmayer University of Osijek

Abstract
This paper considers multiple-circle detection problem on the basis of given data. The problem is being solved by application of center-based clustering method. For the purpose of searching a locally optimal partition, modeled on a well-known $k$-means algorithm, $k$-closest circles algorithm has been constructed. The method has been illustrated with several numerical examples.

Key words: data clustering, circle detection, $k$-means, locally optimal partition

1 Introduction

Grouping a data set into clusters is an important problem in many applications (e.g. facility location problem, pattern recognition, text classification, business, etc.) (see [8, 11, 20]).

In the paper we consider multiple-circle detection problem based on given data. This problem appears in several areas, such as pattern recognition, earthquake investigations, etc (see e.g. [16, 18, 19, 20]). In solving the problem we apply a center-based clustering method. The points are grouped around circles, in such a way that the sum of distances from the data points and appropriate closest circles is minimized.

Let $\mathbb{R}^2$ denote the set of all points in the plane and $\mathbb{R}_+$ denote the set of nonnegative real numbers.

A partition of the data-points set $\mathcal{A} = \{A_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\} \subset \mathbb{R}^2$ into $k$ disjoint subsets $\pi_1, \ldots, \pi_k$, $1 \leq k \leq m$, such that

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \quad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \quad |\pi_j| \geq 1, \quad j = 1, \ldots, k,$$
will be denoted by $\Pi(\mathcal{A}) = \{\pi_1, \ldots, \pi_k\}$, and the elements $\pi_1, \ldots, \pi_k$ of such a partition are called \textit{clusters} in $\mathbb{R}^2$.

Let us assume that all data from the set $\mathcal{A}$ come from some circles that we should reconstruct or detect.

To each cluster $\pi_j \in \Pi$ we associate a corresponding circle-representative $C_j(p_j, q_j, r_j)$ with centre $S_j = (p_j, q_j)$ and radius $r_j$ such that

$$ (p_j^*, q_j^*, r_j^*) = \arg\min_{p, q, r \in \mathbb{R}} \Phi_j(p, q, r), \quad \Phi_j(p, q, r) = \sum_{A_i \in \pi_j} D(C(p, q, r), A_i), \quad (1) $$

where $D(C(p, q, r), A_i)$ represents the distance from the point $A_i$ to the circle $C$. There is a number of recent literature focused on this problem [1, 4, 7, 9].

If we consider in (1) that measure of distance from the circle to the point has the form

$$ D(C, A_i) = |\sqrt{(x_i - p)^2 + (y_i - q)^2} - r|, \quad (2) $$

then we talk about the \textit{total least squares} (TLS) optimality criterion.

If we consider that distance from the circle to the point is euclidean distance

$$ D(C, A_i) = |(x_i - p)^2 + (y_i - q)^2 - r|^2, \quad (3) $$

then we apply the \textit{least absolute deviations} (LAD) optimality criterion ([21]).

If we take that

$$ D(C_j, A_i) = |(x_i - p)^2 + (y_i - q)^2 - r|^2, \quad (4) $$

then we have the so-called algebraic fitting criterion.

If we define an objective function $\mathcal{F}: \mathcal{P}(\mathcal{A}, k) \to \mathbb{R}_+$ on the set of all partitions $\mathcal{P}(\mathcal{A}, k)$ of the set $\mathcal{A}$ containing $k$ clusters in the sense of closest circles $C_1, \ldots, C_k$ by

$$ \mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{A_i \in \pi_j} D(C_j, A_i), \quad (5) $$

then an optimal partition $\Pi^*$ is a partition at which function $\mathcal{F}$ attains its minimum, i.e. $\Pi^* = \arg\min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{F}(\Pi)$.

Conversely, for a given set of circles $C_1, \ldots, C_k$, applying the minimal distance principle, we can define the partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ in the following way:

$$ \pi_j = \{A \in \mathcal{A}: D(C_j, A) < D(C_s, A), \forall s = 1, \ldots, k, \ s \neq j\}, \quad j = 1, \ldots, k. $$
Therefore, the problem (5) of finding an optimal partition of the set \( A \) can be reduced to the following optimization problem

\[
\text{argmin}_{C_1, \ldots, C_k \subset \mathbb{R}^2} \quad F(C_1, \ldots, C_k), \quad F(C_1, \ldots, C_k) = \sum_{i=1}^{m} \min_{j=1, \ldots, k} D(C_j, A_i). \tag{6}
\]

In general, the functional \( F \) is not differentiable and it may have several local and global minima.

In Section 2 we look at the problem of fitting the circle, i.e. finding the optimal parameters of circle on the basis of given data. With regard to the problem of data clustering by circles, in Section 3 we give an algorithm for searching a locally optimal partition by means of \( k \)-closest circles. In Section 4 several illustrative examples are mentioned.

2 Geometric and algebraic circle fits

For the problem (1) of locating a circle on the basis of given a set of \( n \) fixed points \( P_i = (x_i, y_i), \ i = 1, \ldots, n \) in the plane, there exists a number of algorithms and methods. These methods are based on different measures and various criteria for defining a "closest" circle (see [2, 4, 7, 15]).

Let \( C(p, q, r) \) be a circle in the plane with center \( S = (p, q) \) and radius \( r \).

If we apply criterion (2), then we have the following optimization problem

\[
\text{argmin}_{p, q, r \in \mathbb{R}} \quad G_2(p, q, r), \quad G_2(p, q, r) = \sum_{i=1}^{n} |\sqrt{(x_i-p)^2 + (y_i-q)^2} - r|^2, \tag{7}
\]

which is based on minimizing the sum of squared distances from the fitting circle to data points. This is the (total) least squares circle fitting ([5, 9]).

For solving optimization problem (7), one can use various iterative algorithms which are successful (e.g. the Levenberg-Marquardt, Landau algorithm, Späth algorithm [5]).

In addition, one can apply criterion (3) and then be faced with the following optimization problem

\[
\text{argmin}_{p, q, r \in \mathbb{R}} \quad G_1(p, q, r), \quad G_1(p, q, r) = \sum_{i=1}^{n} |\sqrt{(x_i-p)^2 + (y_i-q)^2} - r|, \tag{8}
\]

which is based on minimizing the sum of euclidean distances from the fitting circle to the data points. This problem has also been analysed in literature ([2, 14]) and an exact procedure exists for this problem. However, a heuristic
approach (three points method) has also been suggested because it runs much faster than exact procedures.

On the other hand, one can apply algebraic circle fitting wherein the criterion (4) has been taken into account. Then we get the following optimization problem

$$\arg\min_{p,q,r \in \mathbb{R}} G(p, q, r), \quad G(p, q, r) = \sum_{i=1}^{n} |(x_i - p)^2 + (y_i - q)^2 - r^2|^2.$$  \hspace{1cm} (9)

Algebraic circle fitting are noniterative procedures and they give good results in many applications. There exists simple algebraic fitting - Kása method and several modifications such as Chernov-Oroskov modification, and Pratt circle fitting ([4]).

One can also use another criterion of closeness of a circle to data, the so-called minimax criterion, where one is faced with the optimization problem

$$\arg\min_{p,q,r} g(p, q, r), \quad g(p, q, r) = \max_{i=1,\ldots,n} \{|\sqrt{(x_i - p)^2 + (y_i - q)^2} - r|\}.$$  \hspace{1cm} (10)

There are algorithms for solving this problem, too ([3, 15]).

Methods mentioned above for estimation and for searching for optimal parameters have certain properties and advantages, but also disadvantages ([4, 7, 14]). Particular problems frequently determine a choice of appropriate criterion and of specific methods for circle fitting.

### 3 K closest circles algorithm

As we have mentioned in Introduction, the problem of finding an optimal partition of set of points in the plane, $\mathcal{A} = \{A_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\} \subset \mathbb{R}^2$, into $k$ disjoint subsets that are grouped around circles $C_j(S_j(p_j, q_j), r_j), j = 1, \ldots, k$, can be reduced to the following optimization problem (see (5), (6))

$$\arg\min_{C_1,\ldots,C_k \subset \mathbb{R}^2} F(C_1, \ldots, C_k),$$

where

$$F(C_1, \ldots, C_k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} D(C_j, A_i)$$  \hspace{1cm} (10)

Function $F$ given by (5) and function $F$ given by (10) coincide only at optimal partition.

In general, optimization problem (10) is a nonconvex and nonsmooth optimization problem and it could have several local minima. So, one is faced with the complex problem of finding an optimal solution.
One of the most popular clustering algorithms for searching for a locally optimal partition is $k$-means algorithm ([10, 11]). Analogously to $k$-means algorithm, we construct $k$-closest circles algorithm.

Algorithm 1. ($k$-closest circles (KCC) algorithm)

Step 0: Input $1 \leq k \leq m$, $\mathcal{A} = \{A_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$, $\mu = 0$.

Choose an initial partition $\Pi^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_k^{(0)}\}$ and set $\mu = 0$;

Step 1: Solve optimization problem

$$(p_j^{(\mu)}, q_j^{(\mu)}, r_j^{(\mu)}) = \arg\min_{p, q, r \in \mathbb{R}^3} \Phi_j(p, q, r), \quad j = 1, \ldots, k,$$

$$\Phi_j(p, q, r) = \sum_{A_i \in \pi_j^{(\mu)}} D(C(p, q, r), A_i), \quad \text{and set } C_j^{(\mu)} = (p_j^{(\mu)}, q_j^{(\mu)}, r_j^{(\mu)});$$

Step 2: (Assignment step) Determine new partition (new clusters)

$$\Pi^{(\mu + 1)} = \{\pi_1^{(\mu + 1)}, \ldots, \pi_k^{(\mu + 1)}\}$$

according to minimal distance principle

$$\pi_1^{(\mu + 1)} = \{A_i \in \mathcal{A} : D(C_1^{(\mu)}, A_i) < D(C_l^{(\mu)}, A_i), \forall l = 2, \ldots, k\},$$

$$\pi_j^{(\mu + 1)} = \{A_i \in \mathcal{A} \setminus \bigcup_{s=1}^{j-1} \pi_s^{(\mu + 1)} : D(C_j^{(\mu)}, A_i) < D(C_l^{(\mu)}, A_i), \quad \forall l = j + 1, \ldots, k - 1, j = 2, \ldots, k - 1, \pi_k^{(\mu + 1)} = \mathcal{A} \setminus \bigcup_{s=1}^{k-1} \pi_s^{(\mu + 1)}.$$

Step 3: If $\Pi^{(\mu + 1)} = \Pi^{(\mu)}$, STOP. Otherwise, set $\mu = \mu + 1$ and go to Step 1.

In Step 0 the input data have been introduced and initial partition has been chosen. In Step 1, by solving corresponding optimization problem, for each cluster the corresponding circle-representative has been determined. In Step 2, in order to establish new clusters that are grouped around these circle-representatives, the minimal distance principle has been applied.

The following proposition shows that the proposed algorithm has a decreasing property. It enables us to apply the algorithm as a method for obtaining (locally) optimal partition.

Proposition 1. $K$-closest circles algorithm does not increase the value of the objective function $\mathcal{F}$ defined by (5).
Proof. From the proposed Algorithm 1 we obtain clusters \((\pi_1^{(\mu)}, \ldots, \pi_k^{(\mu)})\) and corresponding circles \((C_1^{(\mu)}, \ldots, C_k^{(\mu)})\). It follows that

\[
\mathcal{F}(C_1^{(\mu)}, \ldots, C_k^{(\mu)}) = \sum_{j=1}^{k} \sum_{A_i \in \pi_j^{(\mu)}} D(C_j^{(\mu)}, A_i)
\]

\[(\text{step 2}) \quad \geq \sum_{j=1}^{k} \sum_{A_i \in \pi_j^{(\mu+1)}} D(C_j^{(\mu)}, A_i)\]

\[(\text{step 1}) \quad \geq \sum_{j=1}^{k} \sum_{A_i \in \pi_j^{(\mu+1)}} D(C_j^{(\mu+1)}, A_i) = \mathcal{F}(C_1^{(\mu+1)}, \ldots, C_k^{(\mu+1)}).\]

\(\square\)

4 Examples

In this section we give a few illustrative examples with various synthetic and empirical data. We suppose that the number of clusters \(k\) is given in advance in all examples. Calculations were done by Mathematica [22].

Example 1. On the basis of three circles given in parametric form \(K_i = S_i + r_i(\cos t, \sin t), t \in [0, 2\pi], i = 1, 2, 3\), the set \(\mathcal{A}\) of 85 random points is generated by using binormal random additive errors with mean vector 0 \(\in \mathbb{R}^2\) and the covariance matrix \(\sigma^2 I\), \(\sigma^2 = 0.1\) (see Fig. 1a). The sum of orthogonal distances (LAD criterion) from these points to corresponding circles \(K_1, K_2, K_3\) is \(F = 5.94991\). On the basis of data-points set \(\mathcal{A}\), circles should be reconstructed. By using initial circles shown in Fig. 1b, KCC algorithm is started. In Fig. 2 the first, third, fifth, seventh, and finally ninth iteration
with corresponding objective function value are shown. In the end, objective function value $F^\star = 5.66076$ and circles are obtained, for which it can be said that they represent a good reconstruction of original circles.

Figure 2: A few steps of $k$-closest circles algorithm ($k = 3$)

![Figure 2](image)

<table>
<thead>
<tr>
<th>Circle</th>
<th>Original circles Center</th>
<th>Original circles Radius</th>
<th>Algebraic fitting criterion Center</th>
<th>Algebraic fitting criterion Radius</th>
<th>LAD criterion Center</th>
<th>LAD criterion Radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(3, 2)</td>
<td>3</td>
<td>(2.653, 1.657)</td>
<td>2.942</td>
<td>(2.654, 1.366)</td>
<td>3.205</td>
</tr>
<tr>
<td>2</td>
<td>(-0.5, 0)</td>
<td>1</td>
<td>(-0.574, 0.086)</td>
<td>1.014</td>
<td>(-0.485, 0.122)</td>
<td>0.984</td>
</tr>
<tr>
<td>3</td>
<td>(0, -3)</td>
<td>2</td>
<td>(0.097, -2.815)</td>
<td>1.924</td>
<td>(0.054, -2.814)</td>
<td>1.948</td>
</tr>
<tr>
<td>4</td>
<td>(2, -2)</td>
<td>1.5</td>
<td>(2.007, -1.652)</td>
<td>1.506</td>
<td>(3.432, 0.898)</td>
<td>1.954</td>
</tr>
</tbody>
</table>

Table 1: Centers and radii of circles

Example 2. On the basis of four circles, the set $A$ of 100 pseudorandom points is generated by adding uniformly distributed pseudorandom errors in interval $[0, 0.2]$ (see Fig. 3a). After that, an initial partition with $k = 4$ clusters is obtained by the built-in function FindClusters (see Fig. 3a). By using KCC-algorithm with algebraic circle fitting - criterion (4) (see Fig. 3b) and orthogonal distances - LAD criterion (3) (see Fig. 3c), we obtained corresponding circles. Centers and radii of original circles, and of circles obtained by KCC-algorithm for two criteria are shown in Table 1. We can say that
obtained circles-centers represent a good reconstruction of original circles and given data (with an exception to the fourth circle of LAD criterion).

![Figure 3: Circles reconstruction](image)

**Example 3.** Similarly as in Example 1 synthetic data are generated by four circles with the same center and different radii. With initial partition shown in Fig. 4a the same KCC-locally optimal partition is obtained (see Fig. 4b) by using algebraic circle fitting - criterion (4) and by using orthogonal distances - LAD criterion (3). It can be seen that obtained circles-centers of these partition represent a good reconstruction of the original circles.

![Figure 4: Circles reconstruction](image)

**Example 4.** Similarly as in Example 1 synthetic data are generated by ten circles, with different centers and same radius. With initial partition shown
in Fig. 5a KCC-locally optimal partition is obtained by using algebraic circle fitting - criterion (4) (see Fig. 5b) and orthogonal distances - LAD criterion (3) (see Fig. 5c). We can say that circles-centers obtained by using algebraic circle fitting represent reconstruction of original circles well enough, which cannot be said for circles-centers obtained by using LAD criterion.

(a) Data and initial partition (b) Algebraic criterion (c) LAD criterion

Figure 5: Circles reconstruction

Remark 1. In the case of synthetic data constructed as in previous examples, one could determine a measure of quality of some partition by using Hausdorff distance between the set of original circles and the set of reconstructed circles (see e.g. [18]).

Example 5. In the paper [18] seismic activity data from a wider area of the Republic of Croatia has been considered, in order to locate the most intense seismic activity in the observed area. It has been shown that the optimal partition with $k = 13$ clusters points out at 13 locations in which the most intense seismic activity in the observed area can be expected. It is interesting to find out the geometric position of circles at which some cluster centers are situated (nine points-centers have been taken into account). By using KCC-algorithm with algebraic circle fitting - criterion (4) and orthogonal distances - LAD criterion (3), we obtained corresponding circles (see Fig. 6a and see Fig. 6b).

5 Conclusions

This paper considers the multiple-circle detection problem on the basis of given data-points set which comes from several circles in the plane. In solving the problem, a center-based clustering method has been applied. Let
us note that numerical experiments show that the proposed KCC-algorithm has similar properties as $k$-means algorithm and it can mainly give a locally optimal partition. If we have a good initial approximation, KCC-algorithm can provide an acceptable solution. In case we do not have a good initial approximation, the algorithm should be restarted with various random initializations, as proposed by [13]. It is assumed that the number of clusters is given in advance.

The problem of determining the appropriate number of clusters in a partition is a specific problem that has not been considered in this paper.

By applying the proposed KCC-algorithm, one can see a certain dependence of results of circles reconstruction on different criteria which are implemented for fitting of circles.

Acknowledgement

The author would like to thank Prof. Scitovski (University of Osijek, Croatia) for his support, helpful comments and useful suggestions.

References


