Croatian Operational Research Review
CRORR **5**(2014), 15–24

Data clustering for circle detection

Tomislav Marošević[1,*]

[1] *Department of Mathematics, Josip Juraj Strossmayer University of Osijek*
Trg Lj. Gaja 6, HR-31 000 Osijek, Croatia
E-mail: ⟨tmarosev@mathos.hr⟩

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

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

Received: July 30, 2013; accepted: January 22, 2014; available online: February 27, 2014

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 [9, 12, 22]).

In this paper we consider a multiple-circle detection problem based on given data. This problem appears in several areas, such as pattern recognition, earthquake investigations, image analysis, machining of round parts, etc. (see e.g. [7, 17, 19, 20, 22]). In solving the problem we apply a center-based clustering method. The points are grouped around circles such that the sum of distances from data points and appropriate closest circles is minimized.

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

A partition of the data-points set $\mathcal{A} = \{A_i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\} \subset \mathbb{R}^2$ 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 partition are called *clusters* in $\mathbb{R}^2$.

Let us assume that all data from the set $\mathcal{A}$ come from some circles that should be reconstructed or detected.

*Corresponding author.

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

$$(p_j^\star, q_j^\star, r_j^\star) = \underset{p,q,r\in\mathbb{R}}{\operatorname{argmin}} \Phi_j(p,q,r), \quad \Phi_j(p,q,r) = \sum_{A_i\in\pi_j} D(C(p,q,r), A_i), \tag{1}$$

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, 8, 10].

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

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

then we talk about the *total least squares* (TLS) optimality criterion.

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

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

then we apply the *least absolute deviations* (LAD) optimality criterion ([23]).

If we take that

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

then we have the so-called algebraic fitting criterion.

If we define an objective function $\mathcal{F}\colon \mathcal{P}(\mathcal{A}, k) \to \mathbb{R}_+$ 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

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

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

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:

$$\pi_j = \{A \in \mathcal{A}\colon D(C_j, A) < D(C_s, A),\ \forall s = 1, \ldots, k,\ s \neq j\}, \quad j = 1, \ldots, k.$$

Therefore, problem (5) of finding an optimal partition of the set $\mathcal{A}$ can be reduced to the following optimization problem

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

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 the 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 for 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 problem (1) of locating a circle on the basis of the given 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, 8, 16]).

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

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

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

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

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

which is based on minimizing the sum of euclidean distances from the fitting circle to data points. This problem has also been analysed in literature ([2, 15]) 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 criterion (4) is taken into account. Then we get the following optimization problem

$$\underset{p,q,r\in\mathbb{R}}{\operatorname{argmin}} G(p, q, r), \quad G(p, q, r) = \sum_{i=1}^{n} |(x_i - p)^2 + (y_i - q)^2 - r^2|^2. \tag{9}$$

Algebraic circle fitting refers to noniterative procedures that give good results in many applications. There exists simple algebraic fitting - Käsa 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 deals with the optimization problem

$$\underset{p,q,r}{\operatorname{argmin}} g(p, q, r), \quad g(p, q, r) = \max_{i=1,\ldots,n} \{|\sqrt{(x_i - p)^2 + (y_i - q)^2} - r|\}.$$

There are algorithms for solving this problem, too ([3, 16]).

The above mentioned methods for estimation and for searching for optimal parameters have certain properties and advantages, but also disadvantages ([4, 8, 15]). Particular problems frequently determine the choice of an appropriate criterion and specific methods for circle fitting.

3. K closest circles algorithm

As we have mentioned in the Introduction, the problem of finding an optimal partition of the set of points in the plane, $\mathcal{A} = \{A_i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\} \subset \mathbb{R}^2$, into k disjoint subsets 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))

$$\underset{C_1,\ldots,C_k \subset \mathbb{R}^2}{\operatorname{argmin}} F(C_1, \ldots, C_k),$$

where

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

Function $\mathcal{F}$ given by (5) and function F given by (10) coincide at optimal partition ([21]).

In general, optimization problem (10) is a nonconvex and nonsmooth optimization problem and it could have several local minima. So, one deals with the complex problem of finding an optimal solution.

One of the most popular clustering algorithms for searching for a locally optimal partition is the k-means algorithm ([11, 12]). Analogously to the k-means algorithm, we construct the *k-closest circles algorithm.*

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

Step 0: Input $1 \leq k \leq m$, $\mathcal{A} = \{A_i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$. *Choose an initial partition* $\Pi^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_k^{(0)}\}$ *and set* $\mu = 0$*;*

Step 1: Solve the optimization problem

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

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

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

$\Pi^{(\mu+1)} = \{\pi_1^{(\mu+1)}, \ldots, \pi_k^{(\mu+1)}\}$ *according to the minimal distance principle*

$$\begin{aligned}
\pi_1^{(\mu+1)} &= \{A_i \in \mathcal{A} \colon D(C_1^{(\mu)}, A_i) < D(C_l^{(\mu)}, A_i),\ \forall l = 2, \ldots, k\},\\
\pi_j^{(\mu+1)} &= \{A_i \in \mathcal{A} \setminus \cup_{s=1}^{j-1} \pi_s^{(\mu+1)} \colon D(C_j^{(\mu)}, A_i) < D(C_l^{(\mu)}, A_i),\\
&\qquad \forall l = j+1, \ldots, k\}, \qquad j = 2, \ldots, k-1,\\
\pi_k^{(\mu+1)} &= \mathcal{A} \setminus \cup_{s=1}^{k-1} \pi_s^{(\mu+1)}.
\end{aligned}$$

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 the initial partition has been chosen. In *Step 1*, by solving the corresponding optimization problem, the corresponding circle-representative has been determined for each cluster. In *Step 2*, in order to establish new clusters 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 a (locally) optimal partition.

Proposition 1. *The 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 the corresponding circles $(C_1^{(\mu)}, \ldots, C_k^{(\mu)})$. It follows that

$$\begin{aligned}
\mathcal{F}(C_1^{(\mu)}, \ldots, C_k^{(\mu)}) &= \sum_{j=1}^{k} \sum_{A_i \in \pi_j^{(\mu)}} D(C_j^{(\mu)}, A_i) \\
(Step\ 2) \quad &\geq \sum_{j=1}^{k} \sum_{A_i \in \pi_j^{(\mu+1)}} D(C_j^{(\mu)}, A_i) \\
(Step\ 1) \quad &\geq \sum_{j=1}^{k} \sum_{A_i \in \pi_j^{(\mu+1)}} D(C_j^{(\mu+1)}, A_i) = \mathcal{F}(C_1^{(\mu+1)}, \ldots, C_k^{(\mu+1)}) .
\end{aligned}$$

□

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* [24].

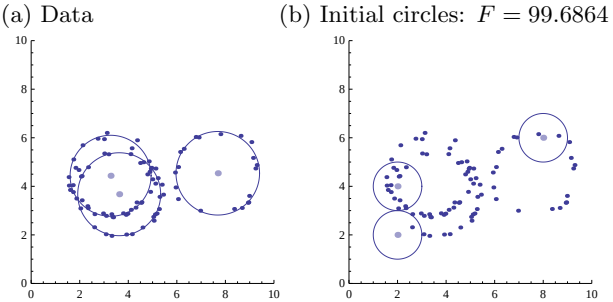

Figure 1: *Initializations for the k-closest circles algorithm ($k = 3$)*

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 $F = 5.94991$. Circles should*

be reconstructed on the basis of the data-points set $\mathcal{A}$. By using the chosen circles shown in Fig. 1b, the initial partition has been determined according to the minimal distance principle. In Fig. 2, the first, the third, the fifth, the seventh, and finally the ninth iteration with the 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.

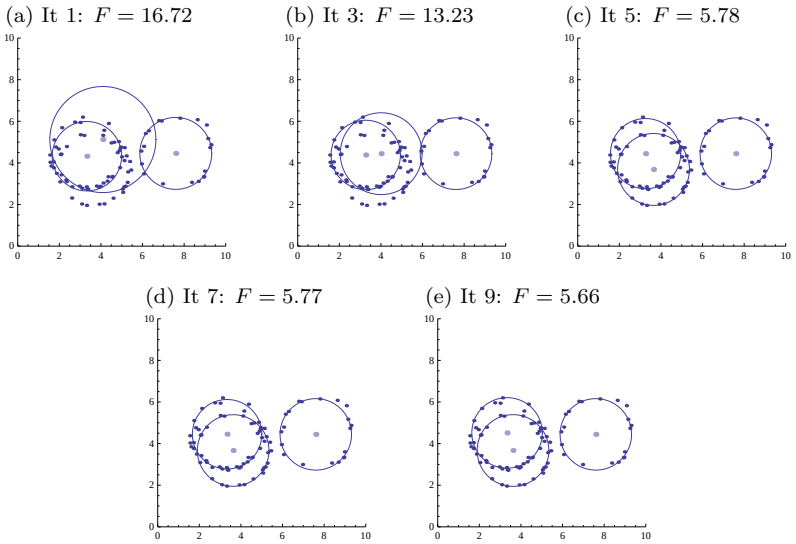

Figure 2: *A few steps of the k-closest circles algorithm ($k = 3$)*

Original 4 circles		Algebraic fitting criterion		LAD criterion		TLS criterion	
Center	Radius	Center	Radius	Center	Radius	Center	Radius
$(3, 2)$	3	(2.65, 1.66)	2.94	(2.65, 1.37)	3.20	(2.66, 1.68)	2.92
$(-0.5, 0)$	1	$(-0.57, 0.09)$	1.01	$(-0.48, 0.12)$	0.98	$(-0.52, 0.19)$	0.95
$(0, -3)$	2	$(0.10, -2.81)$	1.92	$(0.05, -2.81)$	1.95	$(-0.10, -2.77)$	1.91
$(2, -2)$	1.5	$(2.01, -1.65)$	1.51	(3.43, 0.90)	1.95	(1.68, -1.66)	1.83

Table 1: *Centers and radii of circles*

Example 2. *On the basis of four circles, the set $\mathcal{A}$ 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 Mathematica function FindClusters that uses the Euclidean distance and clusters the data based on proximity (see Fig. 3a). By using the KCC-algorithm with algebraic circle fitting - criterion* (4) *(see Fig. 3b), with orthogonal distances - LAD criterion* (3) *(see Fig. 3c) and with TLS criterion* (2)*, we obtained corresponding circles.* Table 1 *shows centers and radii of original four circles and of circles obtained by the KCC-algorithm for three criteria. We can say that the obtained circles-centers*

represent a good reconstruction of original circles and given data (with an exception of the fourth circle of the LAD criterion).

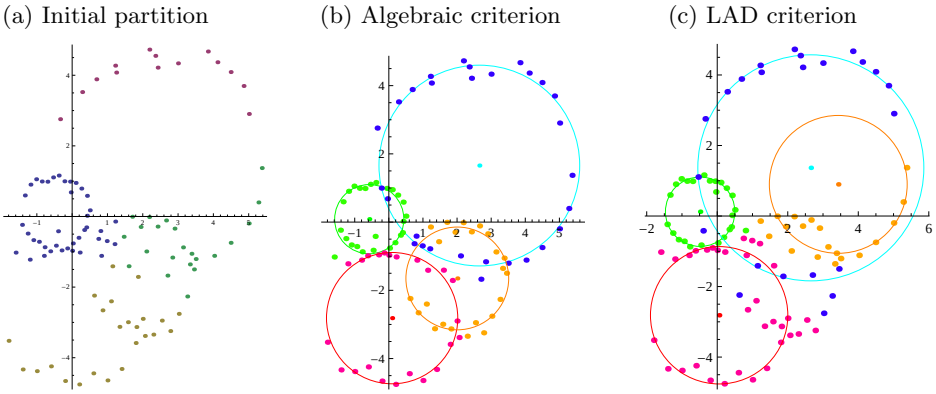

Figure 3: *Circles reconstruction*

Example 3. *Similarly to Example 1, synthetic data are generated by four circles with the same center and different radii. With the initial partition shown in Fig. 4a the same KCC-locally optimal partition is obtained by using algebraic circle fitting - criterion* (4) *(see Fig. 4b), by using orthogonal distances - LAD criterion* (3) *and by using TLS criterion* (2)*. The reconstructed circles by all three criteria* (2), (3) *and* (4) *are almost the same in this case. It can be seen that the obtained circles-centers of this partition represent a good reconstruction of the original circles.*

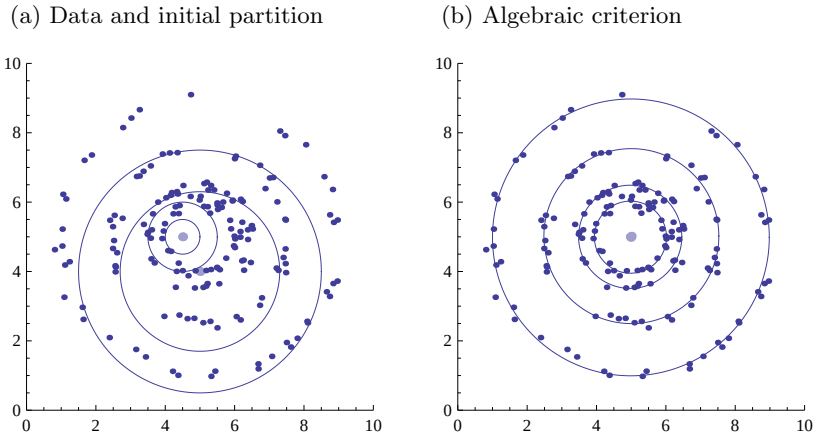

Figure 4: *Circles reconstruction*

Example 4. *Similarly to Example 1, synthetic data are generated by ten circles with different centers and the same radius. With the initial partition shown in Fig. 5a, the 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).*

In this case, TLS criterion (2) *has given similar results as the LAD criterion. We can say that circles-centers obtained by using algebraic circle fitting represent a reconstruction of original circles well enough, which cannot be said for circles-centers obtained by using the LAD criterion.*

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

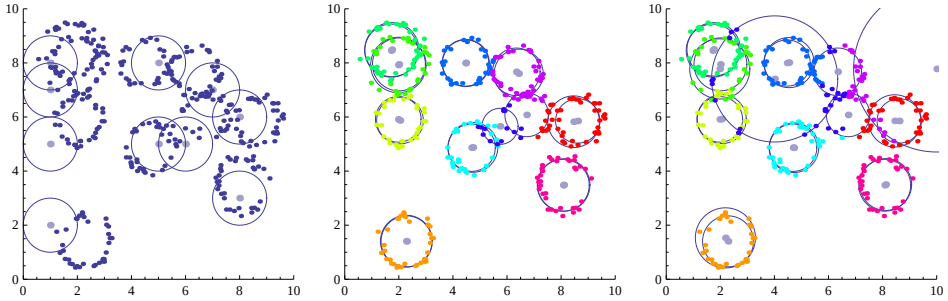

Figure 5: *Circles reconstruction*

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. [19]).*

Example 5. *In paper [19], 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. For the purpose of analyzing the geometric position of circles at which some cluster centers are situated, it is interesting to find out that position (nine points-centers have been taken into account, [19]). Two corresponding*

(a) Algebraic criterion (b) LAD criterion

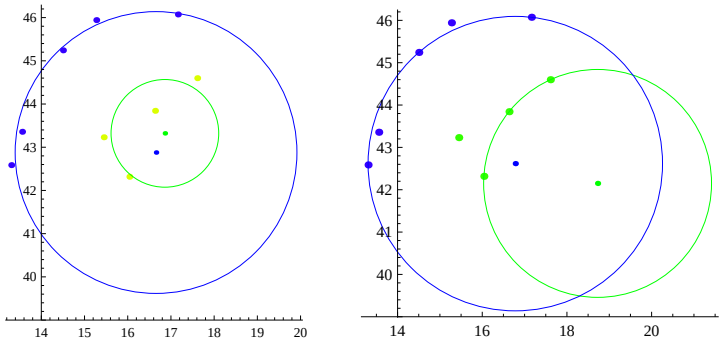

Figure 6: *Circles reconstruction*

circles have been obtained by using the KCC-algorithm: with algebraic circle fitting - criterion (4) *(see Fig. 6a), and with orthogonal distances - LAD criterion* (3) *(see*

Fig. 6b). In this case, TLS criterion (2) *has also given two circles that are by the location of centers and radii between the results of criteria* (4) *and* (3).

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 the k-means algorithm and it can mainly give a locally optimal partition. If we have a good initial approximation, the KCC-algorithm can provide an acceptable solution. In the case we do not have a good initial approximation, the algorithm should be restarted with various random initializations, as proposed by [14]. 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 implemented for fitting of circles. It seems that algebraic criterion (4) has a certain advantage since it includes both the smallest and the largest distance of the point from the circle ($(d^2 - r^2)^2 = (d - r)^2 \cdot (d + r)^2$); however, a more extensive investigation should be done with respect to this matter.

Acknowledgement

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

References

[1] Ahn, S. J., Rauh, W., and Warnecke, H.J. (2001). Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola. Pattern Recognition, 34, 2283-2303.

[2] Brimberg, J., Juel, H., and Schöbel, A. (2009). Locating a minisum circle in the plane. Discrete Applied Mathematics, 157, 901-912.

[3] Brimberg, J., Juel, H., and Schöbel, A. (2006). Locating a circle on the plane using the minimax criterion. IMM Technical Report, No. 1-2006.

[4] Chernov, N. (2010). Circular and linear regression: Fitting circles and lines by least squares. Chapman&Hall/CRC.

[5] Chernov, N., and Lesort, C. (2005). Least Squares Fitting of Circles. Journal of Mathematical Imaging and Vision, 23, 239-252.

[6] Cupec, R., Grbić, R., Sabo, K., and Scitovski, R. (2009). Three points method for searching the best least absolute deviations plane. Applied Mathematics and Computation, 215, 983-994.

[7] Drezner, Z., and Brimberg, J. (2013). Fitting concentric circles to measurements. Mathematical Methods of Operations Research, DOI 10.1007/s00186-013-0455-4, Springer, Published online.

[8] Drezner, Z., Steiner, S., and Wesolowsky, G.O. (2002). On the circle closest to a set of points. Computers&Operations Research, 29, 637-650.
[9] Gan, G., Ma, C, and Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. Philadelphia: SIAM.
[10] Gander, W., Strebel, R., and Golub, G.H. (1995). Fitting of circles and ellipses least squares solution. SVD and Signal Processing III, 349-356.
[11] Jain, J. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31, 651-666.
[12] Kogan, J. (2007). Introduction to Clustering Large and High-Dimensional Data. Cambridge: Cambridge University Press.
[13] Kogan, J., and Teboulle, M. (2006). Scaling clustering algorithms with Bregman distances. In: Berry, M.W., and Castellanos, M. (Eds.). Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining.
[14] Leisch, F. (2006). A toolbox for k-centroids cluster analysis. Computational Statistics & Data Analysis, 51, 526-544.
[15] Nievergelt, Y. (2010). Median spheres: theory, algorithms, applications. Numerische Mathematik, 114, 573-606.
[16] Nievergelt, Y. (2002). A finite algorithm to fit geometrically all midrange lines, circles, planes, spheres, hyperplanes, and hyperspheres. Numerische Mathematik, 91, 257-303.
[17] Qiao, Y., and Ong, S.H. (2004). Connectivity-based multiple-circle fitting. Pattern Recognition, 37, 755-765.
[18] Sabo, K., and Scitovski, R. (2008). The best least absolute deviations line – properties and two efficient methods. ANZIAM Journal, 50, 185-198.
[19] Scitovski, R., and Scitovski, S. (2013). A fast partitioning algorithm and its application to earthquake investigation. Computers and Geosciences, 59, 124-131.
[20] Song, Q., Yang, X., Soh, Y.C., and Wang, Z.M. (2010). An information-theoretic fuzzy C-spherical shells clustering algorithm. Fuzzy Sets and Systems, 161, 1755-1773.
[21] Späth, H. (1983). Cluster-Formation und Analyse. R. Oldenburg Verlag, München.
[22] Teboulle, M. (2007). A unified continuous optimization framework for center-based clustering methods. Journal of Machine Learning Research, 8, 65-102.
[23] Vazler, I., Sabo, K., and Scitovski, R. (2012). Weighted median of the data in solving least absolute deviations problems. Communications in Statistics - Theory and Methods, 41, 1455-1465.
[24] Wolfram, S. (1991). Mathematica - A System for Doing Mathematics by Computer. Addison-Wesley.