Performance Analysis of Subtractive Clustering Algorithm in Determining the Number and Position of Cluster Centers

DOI: https://doi.org/10.47175/rissj.v2i2.241

| Irwandi¹, * | Opim Salim Sitompul² | Rahmat Widia Sembiring³ |

¹Master of Informatics Engineering Study Program, Faculty of Computer Science and Information Technology, University of North Sumatra
²,³Deparment of Computer Science and Information Technology, University of North Sumatra

*ir.whandi@gmail.com

ABSTRACT

The basic concept of the subtractive clustering algorithm is to choose a data point that has the highest density (potential) in a space (variable) as the center of the cluster. The number and position of the cluster centers formed are influenced by the given radius (r) parameter value. If the radius value is very small, it will result in the neglect of potential data points around the center of the cluster. If the value of the radius parameter is too large, it increases the contribution of all potential data points, thereby canceling the effect of cluster density. The number of cluster centers in the subtractive clustering algorithm is determined based on the iteration process in finding data points with the highest number of neighbors. This study uses the clustering partition as a parameter value to determine a data point (candidate cluster center) will be selected to determine the effect of the radius (r) parameter value on the subtractive clustering algorithm in generating clustering. From the experiments that have been carried out on 4 datasets, the results have been obtained, for dataset 1 the highest average value of fuzzy silhouette with a parameter value of radius (r) 0.35 is 0.9088 and the number of clusters 2. While in dataset 2, the average value The highest fuzzy silhouette with a parameter value of radius (r) 0.40 is 0.6742 and the number of clusters 3. While in dataset 3, the average value of the highest fuzzy silhouette with a parameter value of radius (r) 0.50 is 0.7434 and the number of clusters 3. While in the dataset the last is the fourth dataset, the highest fuzzy silhouette average value with a radius (r) parameter value of 0.50 is 0.6630 and the number of clusters 2. This subcractive clustering algorithm is widely applied in the fields of transportation, GIS, big data, control of electric voltages, electrical energy needs, knowing the area of population density to health such as breast cancer diagnosis, which is related to the needs of human life.

KEYWORDS

subtractive clustering; radius; fuzzy silhouette index, human life

INTRODUCTION

Subtractive Clustering Algorithm (Chiu, 1994) is a clustering method modified from Mountain Clustering (Yager & Filev, 1992). In principle, the subtractive clustering algorithm is based on the size of the density of data points (potential) in a space (variable). The data point with the highest potential value will be selected as the center of the cluster. Potential data points within the specified radius around the cluster center will be deducted from their potential value. Then the algorithm will choose another point that has the potential value of the next highest data point to serve as the center of another cluster. This process is repeated until the predetermined criteria are met (Sarin et al. 2019). The subtractive
clustering algorithm method is a simple and fast clustering method and automatically forms the number of clusters. This method is widely implemented in various fields, Liang et al. (2017) breast cancer diagnosis, Wu & Luo (2017) transport, Radionov et al. (2015) controlling electric current voltage, Laksono, H & Hafis, M. (2013). The need for electrical energy, Azizah N. et al. (2019) to find out the area of population density, Polat & Durduran (2011) Geographical Information System (GIS), Pereira et al. (2016) industrial power grid Smart Grid technology, Mubeen et al. (2017) Bigdata.

The subtractive clustering algorithm method is usually used as a preprocessing step in other algorithms to find the number and position of the cluster center (Rezaeian et al. 2017), including: Kokkinos & Margaritis (2018) automatic selection of exemplars points in the affinity propagation algorithm. Yang et al. (2010) automatic selection of the number and position of cluster centers on fuzzy c-means. Rezaeian et al. (2017) using the subtractive clustering algorithm method on the K-means algorithm and the fuzzy c-means algorithm.

However, in its implementation the subtractive clustering algorithm method requires 4 (four) parameters, namely: radius (r), squash factor (q), accept ratio (ε) and reject ratio (ε) (Chiu, 1994). The four parameters are default values (Liang et al. 2017). According to (Sarin et al. 2019), the radius (r) parameter has an important role in optimizing the subtractive clustering algorithm method. So far, the value of the radius parameter has been determined based on “trial and error”. To produce good clustering, the grouping process must be carried out several times with different radius parameter values.

Several studies have been conducted to estimate the value of the radius parameter and the validity of the clustering results in the subtractive clustering algorithm method, including: Shieh et al. (2013) using a genetic algorithm. Sarin et al. (2019) using linear regression. Shieh & Kuo (2011) proposed a new validity index from the combination of compactness and separation to measure the clustering results of the subtractive clustering algorithm method. Shieh (2014) combines compactness, separation and partition index to measure the clustering results of the subtractive clustering algorithm method.

Silhouette index (Rousseeuw, 1987) is a technique to measure clustering quality in crisp clustering which combines compactness and separation results. Campello & Hruschka (2006) proposed the fuzzy silhouette index method to analyze fuzzy clustering. Subbalakshmi et al. (2015) used a fuzzy silhouette index to determine the optimal number of clusters in the fuzzy c-means algorithm using dynamic data.

**LITERATURE REVIEW**

**Subtractive Clustering Algorithm**

Chiu (1994) proposed a subtractive clustering algorithm which is a modification of the mountain clustering algorithm (Yager & Filev, 1992). Subtractive clustering algorithm, determines the data point that has the highest density to the points (surrounding data) as a candidate for the center of the cluster. The data point with the most neighbors will be selected as the center of the cluster. The data point that is the center of the cluster will be reduced in density. Then the algorithm looks for another data point that has the most neighbors to be the center of the next cluster. This process is repeated until all data points are tested.

In practice, the subtractive clustering algorithm requires 4 (four) parameters (Chiu, 1994), namely: radius (r), squash factor (q), accept ratio (ε) and reject ratio (ε). The radius parameter (r) is a vector that will determine how much influence the cluster center has on each data point that is a candidate for the cluster center. The squash factor (q) parameter is used to avoid cluster centers having close densities. The accept ratio (ε) and reject ratio (ε)
parameters are comparison parameters that determine whether or not a data point (candidate cluster center) will be selected as the cluster center.

According to Wu & Luo (2017), the subtractive clustering algorithm includes three main steps: Suppose there are n data points \( \{x_1, x_2, \ldots, x_n\} \) in an M-dimensional space. Assuming the data is normal.

**Step 1:** Calculate the density (potential) of the data points.

\[
P(x_i) = \sum_{k=1}^{n} \exp \left( -\frac{4\|x_i - x_k\|^2}{r^2} \right)
\]

**Step 2:** Revise the potential of each data point

\[
P(x_i) = P(x_i) - P_{cl} \sum_{k=1}^{n} \exp \left( -\frac{4\|x_i - x_{cl}\|^2}{(r \times q^2)} \right)
\]

**Step 3:** In this step, after the density of each data point is revised. Then, look for the data point that has the highest potential to be selected as the center of the second cluster \( x_{c2} \). This process is repeated until a predetermined potential threshold is obtained, namely:

\[
\frac{d_{min}}{r} + \frac{P_k}{P_{cl}} \geq 1
\]

The results of this subtractive clustering algorithm are cluster center matrices \( C \) and sigma \( \sigma \) which will be used to determine the fuzzy membership function parameter values. In this study, the Gauss membership function was used (Shieh, 2014).

**Fuzzy Silhouette Index**

The Silhouette index method introduced by Rousseeuw (1987) is used to measure the quality of the crisp cluster which combines the values of compactness and separation.

\[
a^j_i = \frac{1}{m_j - 1} \sum_{r \neq 1} d(x^j_i, x^j_r)
\]

\[
b^j_i = \frac{1}{m_n} \sum_{r=1} d(x^j_i, x^m_j)
\]

The value range of the silhouette index is -1 to +1. If the silhouette index value is close to 1, it indicates that the data is right in the cluster, if the silhouette index value is 0 or close to 0 then the data position is on the border of the two clusters. Silhouette index value is calculated by (Rousseeuw, 1986).

\[
SI^j_i = \frac{b^j_i - a^j_i}{\max\{a^j_i, b^j_i\}}
\]

Campello et al. (2006) proposed a silhouette index for fuzzy partitioning by including fuzzy membership values in evaluating clusters. The fuzzy partition is validated using a silhouette index by including the fuzzification process. In the fuzzification process, the fuzzy membership matrix is converted into a crisp matrix.
In the fuzzy silhouette index, the average value of the silhouette cluster is calculated using a weighted average. Each data point value is assigned a weighted value based on the reduction in the value of the largest cluster membership in one cluster. Suppose $x_t$ is a data point that has the first and second highest membership values, denoted $u_{pt}$ and $u_{qt}$, then the weight $w_j$ is calculated using the equation:

$$w(x_t) = u_{pt} - u_{qt}$$

While the fuzzy silhouette index is calculated using the equation:

$$F.Sil (x_t) = \frac{\sum_{i=1}^{n} w_i S_t}{\sum_{i=1}^{n} w_i}$$

**RESEARCH METHODS**

In this study, modifications were made to the parameter values of *accept ratio* ($\bar{\varepsilon}$) and *reject ratio* ($\varepsilon$) in the subtractive clustering algorithm. In the standard subtractive clustering algorithm, the *accept ratio* ($\bar{\varepsilon}$) and *reject ratio* ($\varepsilon$) parameter values are used as comparison parameters which determine whether a data point (candidate cluster center) will be selected or not as the cluster center. Meanwhile, this study uses the clustering partition method as a parameter value to determine whether a data point (candidate cluster center) will be selected or not as the cluster center, so that the influence of the radius ($r$) parameter value on the subtractive clustering algorithm in generating *clustering* can be determined.

**System Overview**

![System Overview](image-url)
RESULTS AND DISCUSSION

Research data
This study uses datasets obtained from the UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets.php, including the Iris dataset, Wholesale customers
dataset, Abalone dataset and Banknote dataset.

Discussion
From the experiments that have been carried out on 4 datasets, the results have been
obtained.

| Table 1. Comparative Results of Testing dataset 1 |
|-----------------------------------------------|
| No. | Parameter Radius \((r)\) | SubClust+Partition | Standard SubClust |
|---|-------------------------|-------------------|-------------------|
| 1  | 0.25                    | 0.8753            | 2                 |
| 2  | 0.30                    | 0.5368            | 6                 |
| 3  | 0.35                    | 0.9088            | 2                 |
| 4  | 0.40                    | 0.9032            | 2                 |
| 5  | 0.45                    | 0.8983            | 2                 |
| 6  | 0.50                    | 0.8939            | 2                 |

The highest fuzzy silhouette average value for dataset 1 in the standard subtractive
clustering algorithm with a radius parameter value \((r)\) of 0.50 is 0.7297 and the number of
clusters 4 while with the proposed method, the highest fuzzy silhouette average value with a
radius parameter value \((r)\) 0.35 is 0.9088 and the number of clusters 2.

| Table 2. Comparative Results of Testing dataset 2 |
|-----------------------------------------------|
| No. | Parameter Radius \((r)\) | SubClust+Partition | Standard SubClust |
|---|-------------------------|-------------------|-------------------|
| 1  | 0.25                    | 0.6297            | 3                 |
| 2  | 0.30                    | 0.6484            | 3                 |
| 3  | 0.35                    | 0.6735            | 3                 |
| 4  | 0.40                    | 0.6742            | 3                 |
| 5  | 0.45                    | 0.5585            | 6                 |
| 6  | 0.50                    | 0.5612            | 6                 |

While in dataset 2, the highest fuzzy silhouette average value for dataset 2 in the standard
subtractive clustering algorithm with a radius \((r)\) 0.40 parameter value of 0.6196 and the
number of clusters 2, while with the proposed method, the highest fuzzy silhouette average
value with the parameter value radius \((r)\) 0.40 is 0.6742 and the number of clusters 3.

| Table 3. Comparative Results of Testing dataset 3 |
|-----------------------------------------------|
| No. | Parameter Radius \((r)\) | SubClust+Partition | Standard SubClust |
|---|-------------------------|-------------------|-------------------|
| 1  | 0.25                    | 0.6646            | 2                 |
| 2  | 0.30                    | 0.6968            | 3                 |
| 3  | 0.35                    | 0.7147            | 5                 |
| 4  | 0.40                    | 0.7007            | 5                 |
| 5  | 0.45                    | 0.7171            | 3                 |
| 6  | 0.50                    | 0.7434            | 3                 |
While in dataset 3, the highest average fuzzy silhouette value for dataset 3 in the standard subtractive clustering algorithm with a parameter value of radius \((r)\) 0.50 is 0.7434 and the number of clusters 3, while with the proposed method, the average value of fuzzy silhouette is the highest, with the parameter value radius \((r)\) 0.50 is 0.7434 and the number of clusters 3.

| No. | Parameter | SubClust+Partisi | SubClust Standar |
|-----|-----------|-----------------|------------------|
|     | Radius \((r)\) | Av. FSil | Cluster | Av. FSil | Cluster |
| 1   | 0.25      | 0.5001  | 4     | 0.5902  | 19     |
| 2   | 0.30      | 0.5095  | 10    | 0.5989  | 14     |
| 3   | 0.35      | 0.5254  | 10    | 0.5254  | 10     |
| 4   | 0.40      | 0.5237  | 10    | 0.5030  | 6      |
| 5   | 0.45      | 0.6516  | 2     | 0.4573  | 5      |
| 6   | 0.50      | 0.6630  | 2     | 0.5052  | 4      |

While in the last dataset, namely the fourth dataset, the highest fuzzy silhouette average value for dataset 4 in the standard subtractive clustering algorithm with a radius \((r)\) parameter value of 0.30 is 0.5989 and the number of clusters 14 while with the proposed method, the average value The highest fuzzy silhouette with a parameter value of radius \((r)\) 0.50 is 0.6630 and the number of clusters 2.

**CONCLUSION**

From all the experiments carried out, the value of the radius \((r)\) parameter has not fully guaranteed to increase the fuzzy silhouette value, this is because the subtractive clustering algorithm determining the cluster center point is influenced by four parameter values, namely the radius parameter value \((r)\), the squash factor parameter value \((q)\), accept ratio \((\varepsilon)\) and reject ratio \((\bar{\varepsilon})\).

The test affects four parameter values, namely the radius parameter value \((r)\), the squash factor parameter value \((q)\), accept ratio \((\varepsilon)\) and the reject ratio \((\bar{\varepsilon})\) pembentukan in the formation of clustering in the subtractive clustering algorithm. Comparison or application of other clustering evaluation methods on datasets that have a larger amount of data for better clustering results against clustering results in the subtractive clustering algorithm.

**REFERENCES**

Azizah, N., Yuniarti, D. & Goejantoro, R. (2019). Penerapan Metode Fuzzy Subtractive Clustering. *Jurnal Eksponensial*, [S.l.], v.9, n.2, p.197-206, Available at: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/316

Campello, R.J.G.B. & Hruschka, E.R. (2006). A fuzzy extension of the silhouette with criterion for cluster analysis. *Fuzzy Sets and Systems* 157: 2858-2875.

Chiu, S.L. (1994). Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems* 2: 267-278.

Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. 2nd Edition. Elsevier: San Francisco.

Kokkinos, Y. & Margaritis, K.G. (2018). Kernel averaged gradient descent subtractive clustering for exemplar selection. *Evolving Systems* 9: 285-297.

Laksono, H & Hafis, M. (2013). Aplikasi Fuzzy Clustering dengan Menggunakan Algoritma Subtractive Clustering untuk Perkiraan Kebutuhan Energi Listrik Jangka Panjang di Provinsi Sumatera Barat arri Tahun 2012-2021. *Jurnal Teknologi Informasi &
Liang, M., Huang, L. & Ahmad, W. (2017). Breast cancer intelligent diagnosis based on subtractive clustering adaptive neural fuzzy inference system and information gain. *International Conference on Computer Systems, Electronics and Control (ICCSEC)*, pp. 152-156.

Mubeen, A., Abhinav, N.D. & Swamy, C.V.S. (2017). Reducing the risk of customer migration by using bigdata clustering algorithm. *2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT)*, pp. 992-996.

Pereira, R., Fagundes, A., Melicio, R., Mendes, V.M.F., Figueiredo, J., Martins, J. & Quadrado, J.C. (2016). A fuzzy clustering approach to a demand response model. *Electricity Power and Energy Systems* 81: 184-192.

Polat, K. & Durduran, S.S. (2011). Subtractive clustering attribute weighting (SCAW) to discriminate the traffic accidents on Konya-Affyonkarahisar highway in Turkey with the help of GIS: A case study. *Advances in Engineering Software* 42: 491-500.

Radionov, A.A., Evdokimov, S.A., Sarlybaev, A.A. & Karandaeva, O.I. (2015). Application of subtractive clustering for power transformer fault diagnostics. *International Conference on Industrial Engineering*, pp. 22-28.

Rezaeian, M.H., Esmaeili, S. & Fadaeinedjad, R. (2017). Generator coherency and network partitioning for dynamic equivalencing using subtractive clustering algorithm. *IEEE Systems Journal*: 1-11.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53-65.

Sarin, K., Hodashinsky, I. & Filimonenko, I. (2019). Linear regression to determine the cluster radius for fuzzy rule base generation. *Proceedings of the International Siberian Conference on Control and Communications (SIBCON)*, pp. 1-4.

Shieh, H.-L. (2014) Robust validity index for a modified subtractive clustering algorithm. *Applied Soft Computing* **xxx**: 1-13.

Subbalakshmi, C., Krishna, G.R. Rao, S.K.M. & Rao, P.V. (2015). A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set. *International Conference on Information and Communication Technologies (ICICT 2014)*, pp. 346-353.

Wu, H. & Luo, Q. (2017). A scientific judgment on overload transport by subtractive fuzzy c-means algorithm & three-way decisions. *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017)*, pp.1429-1434.

Yager, R.R. & Filev, D.P. (1992). Generation of fuzzy rules by mountain clustering. *Journal of Intelligent and Fuzzy Systems* 2(3): 209-219.

Yang, Q., Zhang, D. & Tian, F. (2010). An initialization method for fuzzy c-means algorithm using subtractive clustering. *Third International Conference on Intelligent Network and Intelligent Systems*, pp. 393-396.