Application of Subtractive Clustering in Data Processing

Boris Barišić, Aleksandra Krstić, Sanja Vujnović, Željko Đurović

***Abstract*—The problem of data clustering is still in development and various approaches to solving it are being proposed, all of which have different success rates. One of the nonparametric clustering methods is subtractive clustering. The success of this algorithm largely depends on tuning its parameters. In this paper we give a theoretical analysis of different suggestions for choosing their values. Based on probability theory, we examined the impact of dimensionality and number of samples on the clustering radius. By conducting a controlled experiment with known sample distributions, the performance of this algorithm with suggested parameters is tested, as well as its robustness.**

***Index Terms*—subtractive clustering, parameters tuning, classification, data processing**

I. Introduction

For a large number of classification problems we do not have adequate a priori knowledge and therefore we are not able to generate an appropriate training set. Starting with Charles Darwin and his systematization of animals and plants into genera, species, families etc. up until the development of systems based on various forms of artificial intelligence, the man has attempted to improve clustering techniques.

Nonparametric clustering methods do not consider optimization criteria or data distributions. They are based on implementing different ways to locate 'valleys' or 'hills' in the probability density function of the data as a natural border between different classes.

In 1992. Yager and Filev suggested *mountain clustering* as one of the techniques [1]. The method is based on dividing the entire data space into a dense grid of small hypercubes whose vertices are potentially cluster centers. The potential i.e., the *mountain* function is then calculated for each vertex as a measure of sample density in its surrounding. Clearly the potential for the vertex to be a cluster center increases with the number of samples surrounding it. The core idea of this method is the following: after finding the first cluster center, potentials of all vertices are reduced inversely proportional to the distance from the vertex to the cluster center. For vertices closer to the center, the potential reduces more. The next cluster center is chosen as the vertex with the highest potential (after reduction). This method of finding cluster centers repeats until the potential in all vertices falls beneath a certain threshold.

Even though it was imagined as a very simple method, its numerical complexity grows exponentially with sample dimensionality due to a large number of hypercubes in the grid. In 1994. Chiu suggested a modification of this algorithm called subtractive clustering [2].

II. Theoretical analysis

A. Algorithm

The idea of *subtractive* clustering is that every sample in the dataset can be a cluster center. Due to this starting assumption of considering only given samples as cluster centers, the complexity of this algorithm is practically linear.

Samples are assigned a certain density based on which the cluster centers are found during the iterative procedure. Aside from considering a smaller dataset than *mountain* clustering, calculating the density function implies the squared distance between samples, so there is no need for determining the square root.

Based on the description given in [2], the algorithm is comprised of a few steps:

Step 0: For each of N samples we calculate the value of the initial density D_i^1 according to equation (1),

$$D_i^1 = \sum_{j=1}^{N} e^{-\frac{\|X_i - X_j\|^2}{(r_a/2)^2}}, i = \overline{1,N} \tag{1}$$

where r_a is a positive constant called the clustering radius.

Step 1: Based on initial density values we determine the first cluster center X_c^1:

$$X_c^1 = \arg\{\max_{i=1,N} D_i{}^1\} \tag{2}$$

Step 3: Having found the first cluster center, we start the iterative procedure of finding other cluster centers. Since the first center is already found, let the iteration counter start at $k = 1$.

Step 4: We increment the counter to $k = k + 1$ and eliminate the influence of samples near the previously found center by modifying their density function:

$$D_i^k = D_i^{k-1} - D_c^{k-1} e^{-\frac{\|X_i - X_c^{k-1}\|^2}{(r_b/2)^2}}, i = \overline{1,N} \tag{3}$$

Boris Barišić – School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: boris.barisic@ etf.bg.ac.rs).

Aleksandra Krstić – School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: amarjanovic@ etf.bg.ac.rs).

Sanja Vujnović – School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: svujnovic@etf.bg.ac.rs).

Željko Đurović – School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: zdjurovic@etf.bg.ac.rs).

 ISBN 978-86-7466-930-3

D_i^k is the new density function value, and D_i^{k-1} is the previous one. D_c^{k-1} is the maximum density from the previous iteration, and X_c^{k-1} is the cluster center found also in the previous iteration. Parameter r_b represents a new clustering radius.

Step 5: Based on the newly modified values of density functions we choose a k[th] cluster center X_c^k:

$$X_c^k = \arg\{\max_{i=\overline{1,N}} {D_i}^k\} \tag{4}$$

Step 6: We check whether D_c^k i.e., the density of samples in radius r_b of k[th] cluster center satisfies:

$$D_c^k \leq \delta D_c^1 \tag{5}$$

If the given condition is not satisfied, we go back to step 4, and if it is, the algorithm ends. Parameter δ is a positive value smaller than 1 and it is called the clustering threshold.

B. Parameter tuning

The success of clustering largely depends on choosing the values for r_a, r_b and δ. Clustering radiuses r_a and r_b should in some way incorporate the information on how samples are scattered around cluster centers.

A great number of scientists and engineers have put their knowledge and experience into determining clear and straight-forward recommendations for choosing the values of parameters involved in clustering. Unfortunately, most of them so far have turned out to be inefficient and inapplicable in most cases. In [3] it is suggested and in [4] analyzed that the parameter r_a should be chosen according to:

$$r_a = \left\{\frac{1}{4}\left[\max\{\|X_i - X_j\|\} + \min\{\|X_k - X_l\|\}\right]\right\}^{\beta} \tag{6}$$

where $i, j, k, l = \overline{1, N}$, and N the total number of samples available. Parameter β should serve as amortization for extreme values of maxima which come from potential outliers [3].

We will perform an analysis on how the theoretical probability density function (pdf) of r_a changes according to N and dimensionality n. First, we consider a one-dimensional case ($n = 1$). Let samples from all clusters have a Gaussian joined probability density function $\mathcal{N}(m, \sigma^2)$. For the sake of simplicity, we will consider the minimum value to be neglectable in respect to the maximum value, as well as the parameter $\beta = 1$. Samples X_i and X_j then become independent identically distributed random variables, and parameter r_a becomes a random variable R_a for which the following holds:

$$R_a = \max\{|X - Y|\} \tag{7}$$

Random variable $Z = X - Y$ will then also have a Gaussian pdf with parameters $\mathcal{N}(0, 2\sigma^2)$, as it is a subtraction of two Gaussian variables. The absolute value of this variable $U = |Z|$ will have a cumulative distribution function (cdf):

$$F_U(u) = P(U \leq u) = P(|Z| \leq u) = P(-u \leq Z \leq u)u_H(u) \tag{8}$$

i.e. the following will hold:

$$F_U(u) = (2F_Z(u) - 1)u_H(u) \tag{9}$$

where $u_H(u)$ represents a Heaviside unit step function.

Now we can easily obtain the probability density function of the random variable U:

$$f_U(u) = 2f_Z(u)u_H(u) \tag{10}$$

If the total number of samples in our dataset is N, then there are $N_U = \binom{N}{2} = \frac{N(N-1)}{2}$ values which can be calculated as $U = |X - Y|$ and let these values be $U_1, \dots, U_{N_U}$. Let us form an array $U_{(1)}, \dots, U_{(N_U)}$, whose elements are variables $U_1, \dots, U_{N_U}$ in a non-declining order. $U_{(1)}$ is the minimum, and $U_{(N_U)}$ the maximum calculated value. If we want to determine the cdf $F_{U_{(k)}}(u)$ for $U_{(k)}$, we will notice that the event of $\{U_{(k)} \leq u\}$ occurs if and only if the k[th] value is not bigger than u, meaning that at least k of N_U random variables $U_1, \dots, U_{N_U}$ have a value less than or equal to u. Imagine we have N_U Bernoullie's experiments, with every experiment testing whether the event $\{U_{(k)} \leq u\}$ (success) occured or not [5]. The probability of success is equal to $P(U_{(k)} \leq u) = F_U(u)$, and the probability of at least k successes occuring in N_U experiments is:

$$F_{U_{(k)}}(u) = P(U_{(k)} \leq u) = \sum_{i=k}^{N_U} \binom{N_U}{i} F_U^i(u)(1 - F_U(u))^{N_U - i} \tag{11}$$

Since the object of our analysis is the maximum value of $|X - Y|$ i.e. $R_a = U_{(N_U)}$, by replacing k with N_U we get:

$$F_{R_a}(r_a) = \left(F_U(r_a)\right)^{\frac{N(N-1)}{2}} \tag{12}$$

The pdf for the parameter R_a is:

$$f_{R_a}(r_a) = \frac{N(N-1)}{2}\left(F_U(r_a)\right)^{\frac{N(N-1)}{2} - 1} f_U(r_a) \tag{13}$$

Let us assume now that our samples are n-dimensional. Let samples $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ be independent identically distributed vectors whose pdf is $\mathcal{N}(\boldsymbol{M}, \boldsymbol{\Sigma})$. We will additionally assume that their variances along all dimensions are equal and not correlated, i.e. that the covariance matrix has the form $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

The Euclidian norm $V = \|\boldsymbol{X} - \boldsymbol{Y}\|$ is calculated as

$$V = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} = \sqrt{\sum_{i=1}^{n} Z_i^2} \tag{14}$$

where Z_i is the random variable $Z_i = X_i - Y_i$. We previously showed that the distribution of Z_i will be $\mathcal{N}(0, 2\sigma^2)$, which means that $\frac{Z_i}{\sqrt{2}\sigma}$ will have a Gaussian pdf $\mathcal{N}(0,1)$. We can determine the distribution of random variable

 ISBN 978-86-7466-930-3

$$V = \sqrt{2}\sigma \sqrt{\sum_{i=1}^{n} \left(\frac{Z_i}{\sqrt{2}\sigma}\right)^2} \tag{15}$$

using the results known from probability theory, which state that the square root of the sum of squares of n independent identically distributed variables with distribution $\mathcal{N}(0,1)$ will have a χ distribution with n degrees of freedom [6]. Therefore, the cdf of random variable V will be:

$$F_V(v) = \frac{\gamma\left(\frac{n}{2}, \frac{v^2}{4\sigma^2}\right)}{\Gamma\left(\frac{n}{2}\right)} u_H(v) \tag{16}$$

where Γ is the gamma function, and γ is the lower incomplete gamma function.
The pdf of random variable V is:

$$f_V(v) = \frac{v^{n-1}}{2^{n-1}\sigma^n \Gamma\left(\frac{n}{2}\right)} e^{-\frac{v^2}{4\sigma^2}} \cdot u_H(v) \tag{17}$$

Results given in equations (12) and (13) also hold for the distribution of the maximum value of random variable V with arbitrary dimensionality n.

Fig. 1 shows the probability density function of random variable R_a for different dimensionalities n and different number of samples N. We can see that the clustering radius can be affected by changes in both parameters. Larger dimensionality n results in more additions when calculating the norm, which expectedly also results in larger values of clustering radius. On the other hand, having more samples (larger N) increases the chance of extrema appearing, i.e. the chance of having samples which are far away from each other, which has a bigger clustering radius as a result.

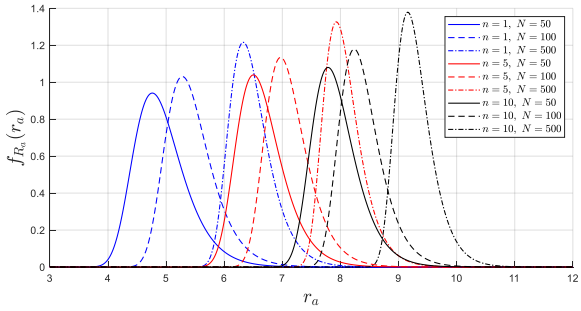

Fig. 1. Probability density function of random variable R_a which represents maximum of Euclidian norm between two samples.

It is also meaningful to analyze how the mathematical expectation m_{R_a} and variance $\sigma^2_{R_a}$ of the pdf of variable R_a change depending on sample number N and dimensionality n. Both statistics have been calculated using numerical integration of pdf in equation (13) and the results are shown in Fig. 2 and Fig. 3. Mathematical expectation and variance increase with the increase of n. On the other hand, larger number of samples increases the mathematical expectation, but decreases variance.

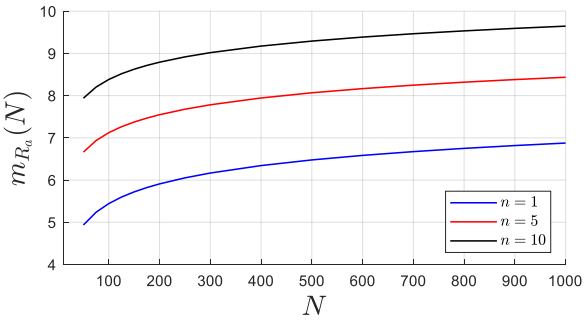

Fig. 2. Mathematical expectation m_{R_a} of random variable R_a depending on dimensionality n and number of samples N.

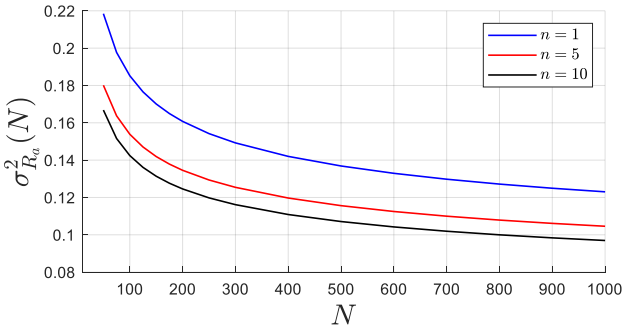

Fig. 3. Variance $\sigma^2_{R_a}$ of random variable R_a depending on dimensionality n and number of samples N.

In the beginning of this analysis, we assumed that parameter $\beta = 1$, which does not have to be the case in general. Let us introduce a random variable $W = R_a^{\beta}$ to show how parameter β affects the clustering radius. We can easily obtain the cdf for random variable W:

$$F_W(w) = P\left(R_a^{\beta} \le w\right) = P\left(R_a \le w^{\frac{1}{\beta}}\right) = F_{R_a}\left(w^{\frac{1}{\beta}}\right) u_H(w) \tag{18}$$

as well as its pdf:

$$f_W(w) = \frac{1}{\beta} w^{\frac{1}{\beta}-1} f_{R_a}\left(w^{\frac{1}{\beta}}\right) u_H(w) \tag{19}$$

For a constant number of samples $N = 100$ and various values of β, the pdf of W is calculated and shown in Fig.4. We can see that by changing the value of β by ± 30 % in respect to the unit value, we make a great impact on the expected value of clustering radius, as well as its variance.

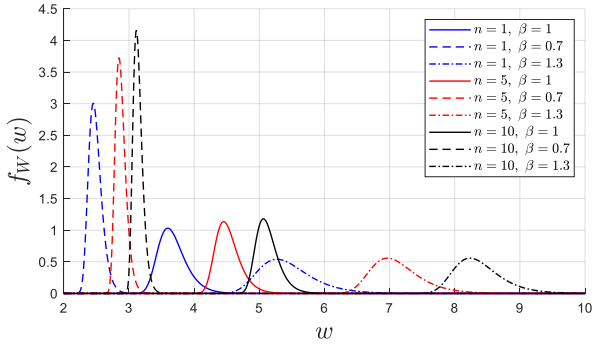

Fig. 4. Probability density function of random variable $W = R_a^{\beta}$ depending on dimensionality of samples n and parameter value β. W represents maximum of Euclidian norm between two samples to the power of β.

 ISBN 978-86-7466-930-3

C. Classification accuracy

Performance of clustering algorithm is tested on two-dimensional samples from 5 classes, distributed normally with different covariance matrices.

The main goal of clustering is to find centers of all classes and then test how accurate the classification is. Since the dataset is synthetically generated and all classes and covariance matrices are known, we decided to determine which class a sample belongs to by calculating the statistical distance according to equation (20),

$$d_i^2 = \left(X - X_{c_i}\right)^T \Sigma_{c_i}^{-1} \left(X - X_{c_i}\right) \tag{20}$$

where d_i is the statistical distance of sample X to i^{th} class, whose center is in X_{c_i}, and covariance matrix is Σ_{c_i}. For each sample the statistical distance to all found centers is calculated, and then the sample gets placed in the class for which the distance is minimal. Classification accuracy is finally measured as a percentage of accurately classified samples.

III. Results and Discussion

We generated 300 samples for each of the 5 classes. As previously mentioned, each class is normally distributed. Clustering radius r_a is determined according to (6), and parameter $\beta = 0.5$ [3]. As suggested in [2], the new clustering radius r_b is $r_b = 1.5r_a$. Clustering threshold δ is manually tuned for the algorithm to detect the right number of clusters.

Figure 6. shows all classes and centers which were detected by subtractive clustering.

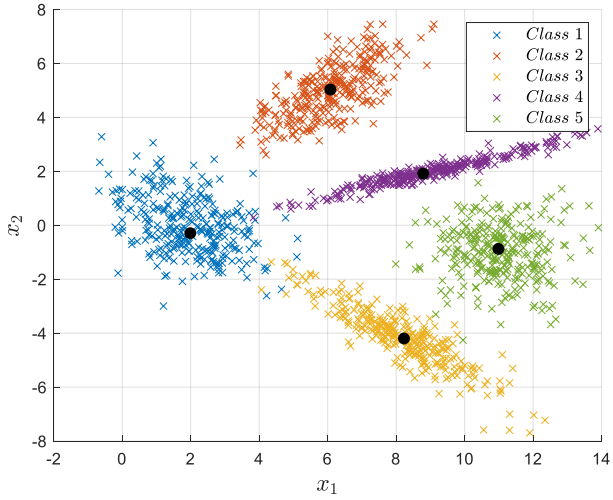

Fig. 5. Sample distribution among classes and corresponding cluster centers found.

To test the sensitivity of the algorithm (with said parameter values) the experiment was repeated 500 times. For each experiment the clustering radius r_a is determined separately. The histogram of the number of detected cluster centers in 500 experiments is given in Fig. 6.

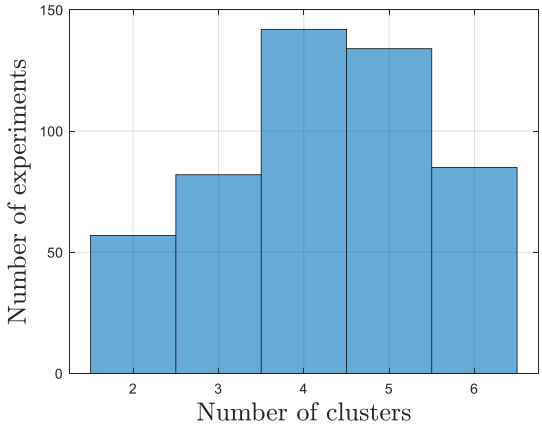

Fig. 6. Histogram of the number of detected cluster centers.

For experiments in which the detected number of centers is correct, i.e. 5, we calculated the classification accuracy as mentioned earlier. Figure 7. shows the histogram of classification accuracy.

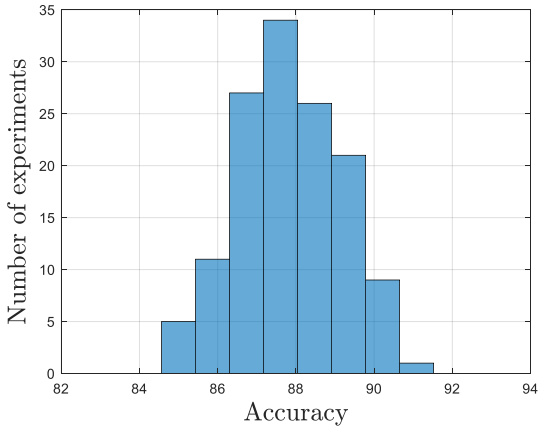

Fig. 7. Histogram of classification accuracy.

In 134 out of 500 experiments the algorithm has managed to detect all cluster centers. The average accuracy in these experiments is 88 %.

Finally, it is also interesting to see how the new cluster radius r_b affects the success of finding cluster centers. We choose o r_b according to:

$$r_b = \varepsilon r_a \tag{21}$$

where ε is a positive constant called the squash factor. Various papers give different suggestions regarding the value of ε, depending on the practical use of the algorithm. Based on [7] and [8] ε should be in an interval of $\varepsilon \in [1, 1.5]$.

We generated the same number of samples with the same distribution as before, and for each of the experiments we tested the success of the algorithm by changing ε from 1 to 1.5 with the step $\Delta\varepsilon = 0.05$. The experiment was repeated one hundered times. Histogram of the number of centers found depending on ε is shown in Fig. 8.

 ISBN 978-86-7466-930-3

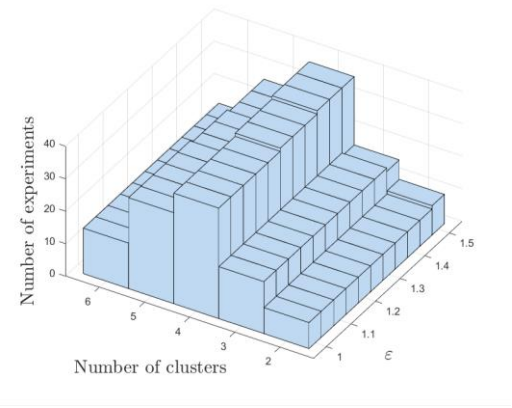

Fig. 8. Histogram of the number of cluster centers found depending on ε.

IV. Conclusion

In this paper we analyzed both theoretically and experimentally the technique of subtractive clustering, as well as suggestions for tuning its parameters. Values of parameters r_a, r_b and δ significantly affect the success of clustering. We observed that even with various suggestions on how to choose the values, the algorithm does not always perform well on different sets with the same distribution.

However, even if we assume that parameters r_a and r_b are tuned correctly, there are no theoretical propositions on choosing the threshold value δ, which has a great impact on the number of clusters found. In case of samples with dimensions $n = 1, 2, 3$ we can visualize the dataset and assess whether the cluster centers have been found correctly. Nevertheless, in most cases the dimensions will be significantly bigger and there would be no unique way to rate the success of the algorithm.

One of the methods to further enhance the algorithm and propose a uniform way for choosing parameter values would be to separately consider the maximum scattering along each of the axis. For example, if we have two-dimensional samples where the variance along one axis is much larger than the other and classes are near to each other, determining the density in the radius as the maximum quadratic norm of two samples, would lead to poor results.

Though the simplicity of subtractive clustering is its big advantage, the results are not ideal. That being said, it can be used as a preprocessing technique for other more sophisticated methods.

Acknowledgment

The paper was co-funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia. This year's contract number is 451-03-68/2022-14/200103.

References

[1] R. R. Yager and D. P. Filev. "Approximate clustering via the mountain method". *IEEE Transactions on systems, man, and Cybernetics*: pp. 1279-1284, 1994.

[2] S. L. Chiu, "Fuzzy model identification based on cluster estimation", *Journal of Intelligent & fuzzy systems*, vol. 2, no. 3, pp. 267-278, 1994.

[3] X. Cui, S. Liu, and L. Jia, "An improved method of semantic driven subtractive clustering algorithm", in *2015 IEEE 5th International Conference on Electronics Information and Emergency Communication*, pp. 232-235, IEEE, 2015.

[4] M. M. Milošević, "Speaker identification in conditions of emotional speech", Ph.D. dissertation, Signals and Systems, University of Belgrade, Belgrade, Serbia, 2020.

[5] M. Merkle, "Random variabiles", *Probability and statistics for engineers and engineering students* [in Serbian], Belgrade, Serbia, Akademska misao, 2016, ch. 3, sec. 3.6, pp. 81–82.

[6] M. Merkle, "Some important distributions", *Probability and statistics for engineers and engineering students* [in Serbian], Belgrade, Serbia, Akademska misao, 2016, ch. 7, sec. 7.2, pp. 157–159.

[7] K. Demirli, S. Cheng, and P. Muthukumaran, "Subtractive clustering based modeling of job sequencing with parametric search", *Fuzzy Sets and Systems*, vol. 137, no. 2, pp. 235-270, 2003.

[8] X.-x. Jing, L. Zhan, H. Zhao, and P. Zhou, "Speaker recognition system using the improved gmm-based clustering algorithm", in 2010 International Conference on Intelligent Computing and Integrated Systems, pp. 482-485, IEEE, 2010.

 ISBN 978-86-7466-930-3