Classification of sinusitis using kernel three-way c-means

S Hartini¹, Z Rustam¹, J Pandelaki², M Prasetyo², and R E Yunus²

¹Department of Mathematics, University of Indonesia, Depok 16424, Indonesia
²Department of Radiology, Cipto Mangunkusumo Hospital, Jakarta 10430, Indonesia

*e-mail: sri.hartini@sci.ui.ac.id

Abstract. Sinusitis can be defined as acute and chronic sinusitis, according to the duration of symptoms. In this study, kernel three-way c-means, as the modification of the three-way c-means method that used kernel distance instead of Euclidean distance, was used. Three-way c-means itself is the upgrade version of the rough k-means algorithm that integrates three-way weight and three-way assignments to assign data points into clusters with the appropriate weight. The performance was later compared using the sinusitis dataset taken from Cipto Mangunkusumo Hospital, Indonesia, which was consists of 102 acute and 98 chronic sinusitis samples. From the experiments, three-way c-means was obtained 62.09% accuracy, 55.21% sensitivity, 62.76% precision, 68.77% specificity, and 58.59% F1-Score in 1.82 seconds. Meanwhile, kernel three-way c-means with the 8th polynomial kernel was provided 67.48% accuracy, 74.82% sensitivity, 64.52% precision, 60.77% specificity, and 69.12% F1-Score in 2.24 seconds. Therefore, it was concluded that kernel three-ways c-means performs better with the slower running time than the three-way c-means.

Keywords: Sinusitis, Kernel three-way c-means

1. Introduction
Sinusitis is the inflammation of paranasal sinuses that can be classified as acute and chronic according to the clinical findings. Acute sinusitis was recognized by the existence of nasal or postnasal discharge with a daytime cough that occurs for three or four consecutive days. Meanwhile, chronic sinusitis has persistent symptoms, including cough, rhinorrhea, or nasal obstruction, lasting more than 90 days [1]. Both of these diseases are common conditions worldwide, with over 24 million cases in the United States [2] and cause more than 107 million sufferers in China [3].

The diagnosis of sinusitis has several problems. Symptoms of acute sinusitis often overlap with other diagnoses such as allergic rhinitis and the common cold, while there are no specific symptoms to diagnose chronic sinusitis [4]. Meanwhile, accurate diagnosis is essential because different disease needs different treatment.

The diagnosis usually was obtained by analyzing the clinical signs and symptoms, nasal endoscopy, and CT-scan results. Many methods were developed in order to obtain a more accurate diagnosis. In recent studies, fuzzy kernel c-means [5] and kernel spherical k-means [6] were used to distinguish acute and chronic sinusitis. Both of these researches encourage this study to observe the performance of another clustering method in providing the diagnosis of sinusitis. A three-way c-means [7] algorithm is the clustering method that has proved its capability with excellent performance. Therefore, in this study, kernel three-way c-means was proposed and compared with the three-way c-means performance.
2. Material and Methods

2.1. Material
The sinusitis dataset consists of 102 acute and 98 chronic sinusitis. This dataset was obtained from Cipto Mangunkusumo (RSCM) Hospital in Indonesia, and each instance was described by four features, namely age, gender, Hounsfield Unit (HU) scale, and air cavity.

2.2. Methods

2.2.1. Kernel function. There are several kernel functions such as Gaussian Radial Basis Function (RBF), polynomial, and sigmoid kernel function that commonly used. The kernel function was expected to improve the performance of machine learning algorithm when the dataset cannot merely be separated using a linear function. In this study, the polynomial kernel function with the first ten polynomial degrees was used. For every \( x, y \in \mathbb{R}^n \), the polynomial kernel function with the polynomial degree \( h \) was defined as below [8].

\[
    K(x, y) = (x \cdot y + 1)^h
\]

Hereafter, the formula above was substituted to the distance definition between two points \( x \) and \( y \) in the feature space [9] that was shown in Eq. (2). This equation then used to replace the Euclidean distance in the algorithm.

\[
    d(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}
\]

2.2.2. Three-way c-means. This method was proposed by Zhang K [7] to fix the problem in rough k-means clustering. This method used three regions of clusters, namely positive region, boundary region, and negative region. The positive region \( \mathbb{C} \) of a cluster was filled by the data points that belong to the cluster, the boundary region \( \hat{\mathbb{C}} \) of a cluster was filled by the data points that probably belong to the cluster, and the data points that do not belong to the cluster are taken as a member of the negative region \( \breve{\mathbb{C}} \) of a cluster. Then, a cluster \( \mathbb{C}_i \) is represented as \( (\mathbb{C}_i, \hat{\mathbb{C}}_i, \breve{\mathbb{C}}_i) \).

Three-way c-means integrates the three-way weight and three-way assignment. For calculating the three-way weight \( w_{ij} \) of data point \( x_j \) in cluster \( \mathbb{C}_i \), the values of \( B_{x_j}, \mu_{ij} \), and \( M_{x_j} \) have to found first. \( B_{x_j} \) is a set of clusters whose positive or boundary regions comprise \( x_j \). Then the fuzzy membership \( \mu_{ij} \) is defined by Bezdek [10] in Eq. (3). Meanwhile \( M_{x_j} \) is a multiset which its elements are the fuzzy membership of \( x_j \) in the clusters of \( B_{x_j} \).

\[
    \mu_{ij} = \sum_{l=1}^{c} \left( \frac{d(x_i, x_l)}{d(x_l, x_i)} \right)^{\frac{2}{m-1}}, \quad 1 \leq i \leq c, \quad 1 \leq j \leq n
\]

This method consists of two steps: update step and assignment step. Update step was used to calculate the means; meanwhile, the assignment step was followed the three-way assignment rules. For the simplification, the algorithm of three-way c-means was given in Figure 1 [7]. This algorithm converges if positive and boundary regions of the current iteration keep unchanged with that of the previous iteration, or the number of iterations reaches the maximum number of iterations.

2.2.3. Kernel three-way c-means. This method is the modification of the three-way c-means algorithm that replaces the distance measurement using the Eq. (2). Therefore, the difference between both methods only occurs in fuzzy membership calculation. Therefore, the algorithm of kernel three-way c-means follows the algorithm in Figure 1 with the fuzzy membership \( \mu_{ij} \) was defined as follows.

\[
    \mu_{ij} = \sum_{l=1}^{c} \left( \frac{K(x_i, x_j) - 2K(x_j, x_j) + K(x_j, x_j)}{K(x_l, x_l) - 2K(x_l, x_j) + K(x_j, x_j)} \right)^{-\frac{1}{m-1}}, \quad 1 \leq i \leq c, \quad 1 \leq j \leq n
\]
2.2.4. Performance measure. The performance of three-way c-means and kernel three-way c-means was examined using the confusion matrix to calculate the accuracy, sensitivity, precision, specificity, and F1-Score, which the formulas are given in Eqs. (5)-(9).

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)
\]
\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad (6)
\]
\[
\text{Precision} = \frac{TP}{TP + FP} \quad (7)
\]
\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (8)
\]
\[
\text{F1-Score} = \frac{2 \ast \text{sensitivity} \ast \text{precision}}{\text{sensitivity} + \text{precision}} \quad (9)
\]

where TP and TN are the numbers of chronic and acute sinusitis samples correctly diagnosed, respectively; meanwhile, FN and FP are the numbers of chronic and acute sinusitis samples incorrectly diagnosed, respectively.

3. Results and Discussions
The performance of each method was evaluated according to the percentage of training data that start from 10 to 90 percent. Each training data percentage was used to find the centroid of the clusters using three-way c-means and kernel three-way c-means, respectively. After that, the distance between each point in the testing data and every centroid was calculated. As a result, the cluster of every data point can be determined based on the closest distance that was calculated before.

The sinusitis dataset was first applied to the three-way c-means algorithm, and the performance of this method is given in Table 1. From this table, the best performance was delivered when using 80 percent of the training data. However, if the average of each performance is calculated, we obtain that three-way c-means performance on the sinusitis dataset was 62.09% accuracy, 55.21% sensitivity, 62.76% precision, 68.77% specificity, and 58.59% F1-Score in 1.82 seconds.
Table 1. The performance of three-way c-means on sinusitis dataset

| Percentage of Training Data | Accuracy | Sensitivity | Precision | Specificity | F1-Score | Running Time |
|-----------------------------|----------|-------------|-----------|-------------|----------|--------------|
| 10                          | 63.89    | **65.91**   | 62.37     | 61.96       | **64.09**| 0.61         |
| 20                          | 62.50    | 56.41       | 62.86     | 68.29       | 59.46    | 0.86         |
| 30                          | 62.86    | 55.38       | 61.02     | 69.33       | 58.06    | 1.24         |
| 40                          | 63.33    | 53.45       | 64.58     | 72.58       | 58.49    | 1.48         |
| 50                          | 60.00    | 51.06       | 58.54     | 67.92       | 54.55    | 1.79         |
| 60                          | 61.25    | 51.35       | 59.38     | 69.77       | 55.07    | 2.10         |
| 70                          | **65.00**| 60.71       | 62.96     | 68.75       | 61.82    | 2.47         |
| 80                          | **65.00**| 57.14       | **70.59** | **73.68**   | 63.16    | 2.76         |
| 90                          | 55.00    | 45.45       | 62.50     | 66.67       | 52.63    | 3.09         |

After that, kernel three-way c-means using polynomial kernel function was applied on the sinusitis dataset, and its performance was shown in Table 2. It was concluded that eight polynomial kernel function was the highest in accuracy, sensitivity, precision, and F1-Score according to this table.

Table 2. The performance of kernel three-way c-means using polynomial kernel on sinusitis dataset

| Polynomial Degree | Average of Accuracy | Average of Sensitivity | Average of Precision | Average of Specificity | Average of F1-Score | Average of Running Time |
|-------------------|---------------------|------------------------|----------------------|------------------------|---------------------|-------------------------|
| 1                 | 61.11               | 51.99                  | 62.19                | 69.99                  | 56.45               | 2.33                    |
| 2                 | 62.01               | 55.26                  | 62.62                | 68.58                  | 58.58               | 2.15                    |
| 3                 | 63.14               | 58.61                  | 63.36                | 67.71                  | 60.71               | 2.19                    |
| 4                 | 64.02               | 60.28                  | 64.00                | 67.77                  | 61.91               | **2.04**                |
| 5                 | 65.21               | 63.53                  | 64.51                | 66.78                  | 63.94               | 2.19                    |
| 6                 | 65.46               | 64.84                  | 64.51                | 66.04                  | 64.60               | 2.08                    |
| 7                 | 52.84               | 54.68                  | 51.32                | 50.89                  | 52.69               | 2.16                    |
| 8                 | **67.48**           | **74.82**              | **64.52**            | **60.77**              | **69.12**           | 2.24                    |
| 9                 | 66.75               | 72.68                  | 63.98                | **60.85**              | 67.95               | 2.27                    |
| 10                | 47.28               | 39.59                  | 43.60                | 54.28                  | 41.44               | 2.17                    |

Therefore, the detail performance of kernel three-way c-means was given in Table 3. From this table, the best performance was also yielded when using 80 percent of the training data.

Table 3. The performance of kernel three-way c-means using 8th polynomial kernel on sinusitis dataset

| Percentage of Training Data | Accuracy | Sensitivity | Precision | Specificity | F1-Score | Running Time |
|-----------------------------|----------|-------------|-----------|-------------|----------|--------------|
| 10                          | 64.44    | 68.18       | 62.50     | 60.87       | 65.22    | **0.42**     |
| 20                          | 66.88    | 78.21       | 62.89     | 56.10       | 69.71    | 0.76         |
| 30                          | 67.14    | 76.92       | 61.73     | 58.67       | 68.49    | 1.44         |
| 40                          | 68.33    | 75.86       | 64.71     | 61.29       | 69.84    | 1.52         |
| 50                          | 68.00    | 76.60       | 63.16     | 60.38       | 69.23    | 1.98         |
| 60                          | 67.50    | 75.68       | 62.22     | 60.47       | 68.29    | 3.10         |
| 70                          | **70.00**| **82.14**   | 63.89     | 59.38       | 71.88    | 3.34         |
| 80                          | **70.00**| 76.19       | 69.57     | **63.16**   | **72.73**| 3.75         |
| 90                          | 65.00    | 63.64       | **70.00** | 66.67       | 66.67    | 3.90         |
The use of kernel function in three-way c-means seems to improve the performance measure. According to Table 2 and Table 3, it was concluded that kernel three-way c-means gives better performance than three-way c-means on the sinusitis dataset. The comparison between three-way c-means and kernel three-way c-means using the eighth polynomial kernel can also be seen in Figure 2.

From this figure, we can see that the accuracy, sensitivity, and F1-Score of three-way c-means tend to decrease as the increasing of training data percentage. Meanwhile, the trend of kernel three-way c-means line performance was increasing for almost performance measurement, including the running time. However, the specificity of three-way c-means was much better than when the algorithm using kernel function. Therefore, it can be assumed that kernel three-way was better to predict the chronic sinusitis, and three-way c-means was better in predicting the acute sinusitis.

**Figure 2.** The performance comparison between three-way c-means (blue) and kernel three-way c-means (red) using eighth polynomial kernel on sinusitis dataset
4. Conclusion
Sinusitis can be classified as acute and chronic sinusitis, according to the duration of symptoms and its clinical findings. Accurate and precise diagnosis of sinusitis is essential because it determines the treatment that the patients would obtain. In this study, the sinusitis dataset was obtained from Cipto Mangunkusumo Hospital in Indonesia. This dataset consists of 102 acute and 98 chronic sinusitis samples with age, gender, Hounsfield Unit (HU) scale, and air cavity as the features of each sample.

Kernel three-way c-means was proposed in this study. Then its performance was compared to three-way c-means in accuracy, sensitivity, precision, specificity, F1-Score, and the running time. From the experiments, three-way c-means was obtained 62.09% accuracy, 55.21% sensitivity, 62.76% precision, 68.77% specificity, and 58.59% F1-Score in 1.82 seconds. Meanwhile, kernel three-way c-means with the 8th polynomial kernel was provided 67.48% accuracy, 74.82% sensitivity, 64.52% precision, 60.77% specificity, and 69.12% F1-Score in 2.24 seconds. Therefore, it was concluded that kernel three-ways c-means performs better with the slower running time than the three-way c-means.

As future work, another process of pre-processing data can be used to improve the performance. Other kernel functions and evaluations can also be considered. The variation of the polynomial degree used in polynomial kernel function in kernel three-way c-means can also be the side that interesting to be analyzed.

Acknowledgment
This research was supported financially by the Ministry of Research, Technology, and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with PTUPT 2020 research grant scheme, and the sinusitis dataset was provided by the Department of Radiology, Cipto Mangunkusumo Hospital Indonesia. Therefore, the authors are thankful to both parties for their kindness in supporting this research.

References
[1] American Academy of Pediatrics 2001 Pediatrics 108(3) 798-808
[2] Mustafa M, Patawari P, Ifthikhar HM, Shimmi SC, Hussain SS, and Sien MM 2015 Int. J. Pharm. Sci. Invent. 4(2) 30-36
[3] Shi JB, Fu QL, Zhang H, Cheng L, Wang YJ, Zhu DD, Lv W, Liu SX, Li PZ, Ou CQ, and Xu G 2015 Allergy 70(5) 533-539
[4] Leung RS and Katial R 2008 Prim. Care Clin. Office Pract. 35 11-24
[5] Putri RA, Rustam Z, and Pandelaki J 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052060
[6] Arfiani, Rustam Z, Pandelaki J, and Siahaan A 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052011
[7] Ziang K 2019 Appl. Soft Comput. J. 82 105536
[8] Liu L, Shen B, and Wang X 2014 Research on kernel function of support vector machine Advanced Technology Embedded and Multimedia for Human-centric Computing Lecture (Notes in Electrical Engineering) Eds Huang Y M, Chao H C, Deng D J, Park J (Springer, Dordrecht)
[9] Bishop CM 2006 Pattern recognition and machine learning (New York: Springer)
[10] Bezdek JC 1981 Pattern recognition with fuzzy objective function algorithms (New York: Plenum Press) pp 203–239.