Fuzzy C-Means Clustering with Minkowski and Euclidean Distance for Cerebral Infarction Classification

R Khairi¹, S G Fitri¹, Z Rustam¹*, J Pandelaki²

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

*rustam@ui.ac.id

Abstract. Cerebral infarction is a condition in which the death of neuronal cells, glial cells and blood vessel system is caused by a lack of oxygen and nutrients. The cause of nerve damage is hypoxia, which is a decrease in oxygen pressure in the alveoli which can cause hypoxemia in brain tissue. Cerebral infarction can also be caused by obstruction of blood flow to the brain so that the brain does not get enough oxygen. This situation is called ischemia. The initial stage of ischemic neurons is characterized by the formation of micro vacuolization, which is characterized by cell size that is still normal or slightly reduced, vacuoles occur in the perikaryon area, which can be found in neurons in the hippocampus and cortical 5-15 minutes after hypoxia. The final sign of cell damage due to ischemia is the nucleus which becomes pyknotic and fragmented. To diagnose the presence or absence of cerebral infarction in the brain it is not enough just to use a CT scan, therefore machine learning will also be used to diagnose the presence or absence of cerebral infarction in the brain. For this reason, the authors propose Fuzzy C-Means Clustering with Minkowski and Euclidean Distance as a classification method that has good accuracy, good precision, good memory, and a good F1-score in calcifying patients whose brains experience infarction or not. In this proposed method, Fuzzy C-Means Clustering with Minkowski and Euclidean Distance is a modification of the Fuzzy C-Means Clustering Algorithm. This modification is proposed to increase the detection capacity of Fuzzy C-Means Clustering. The parameterized Minkowski distance metric is adjusted for implementation with FCM with various settings. The experimental results show that this method can improve the results of the FCM grouping with an accuracy of around 88%.

Keywords: Fuzzy C-Means Clustering, Minkowski Distance and Euclidean Distance.

1. Introduction

Stroke is a cerebrovascular disease that is often found in developed countries, such as Indonesia [1]. Stroke is a serious type of disease. About 30% of stroke patients die within 3 months, and 20% of stroke patients are more severe, and 50% of stroke patients can restore their self-care abilities. Where factors that influence recovery depend on the severity of brain damage. Stroke is an abrupt onset of a focal neurological deficit secondary to a vascular event lasting more than 24 hours. An acute stroke refers to the first 24-hour-period of a stroke event [2].
Stroke is caused by several factors, that is hypertension, smoking, unhealthy diet, lack of physical activity, high blood pressure, increased blood sugar and increased blood lipid profile [1]. Hypertension is the most important risk factor, which can increase the risk of stroke by about two to four times [3]. Hypertension or high blood pressure is an increase in systolic blood that is more than 140 mmHg and diastolic blood pressure that is more than 90 mmHg in 2 measurements in an interval of five minutes in a calm state. Causes of hypertension include age, sex, family history, genetic factors, smoking habits of alcoholic beverages, obesity, lack of physical activity and stress [4].

Stroke is classified into two types, namely hemorrhagic stroke and non-hemorrhagic (ischemic) stroke [5]. Hemorrhagic stroke is a stroke caused by weak blood exploding and bleeding to the brain and its surroundings which causes damage to brain cells so that it cannot work properly [6]. Whereas ischemic stroke is a stroke caused by thrombotic or thromboembolic blockages in the arteries [7]. This blockage of the arteries to the brain disrupts blood flow to the brain so that brain cells cannot make enough energy and will stop working [8]. In ischemic stroke, cerebral infarction is a more common condition.

Acute ischemic infarction correlates with the severity of the stroke as well as the functional outcome of all ischemic stroke subtypes. In patients with lacunar stroke, the size of the infarction and the location of the infarct become the differentiator of this subtype from other forms. Infarct size is typically reported only by maximum lesion diameter on axial imaging, which may inadequately characterize the actual volume [9].

About 40% of acute ischemic strokes are caused by proximal intracranial large vessel occlusion (LVO). After arterial occlusion, there can be a temporal growth of the ischemic core into the accumulating area that is modulated by collateral blood flow, the key element setting the pace of the ischemic process. However, most stroke patients with LVO remain untreated because they present beyond the conventional time windows for acute reperfusion therapies.

Treatment for stroke usually starts with hospital treatment. Treatment of stroke can also be done through antihypertensive therapy. Antihypertensive administration is given with consideration not only to the brain but also to other damage, such as the heart and kidneys. Besides, strokes need to be avoided, also need to modify the lifestyle of not smoking, not drinking alcohol and not using cocaine [3].

By doing a CT scan of the brain, the presence or absence of infarction in the patient's brain can be observed [10]. But, to diagnose cerebral infarction in a person's brain, it is not enough just to have a CT scan, but it also requires machine learning. Machine learning can be used to classify various diseases, such as acute sinusitis, diabetes, and colon cancer [11,12,13]. However, in this study, we will classify infarction in the brain.

In this paper, the authors propose the Fuzzy C-Means Clustering method with Minkowski and Euclidean Distance as a classification method that has good accuracy, good precision, good memory, and good F1-scores in calcifying a patient in his brain. cerebral infarction or not. In this proposed method, Fuzzy C-Means Clustering with Minkowski and Euclidean Distance is a modification of the Fuzzy C-Means Clustering Algorithm. This modification is proposed to increase the detection capacity of Fuzzy C-Means Clustering. The parameterized Minkowski distance metric is adapted for implementation with FCM with various settings. The experimental results show that this method can improve the results of FCM grouping [5].

2. Method

2.1. Fuzzy C-Means Clustering

Clustering has an important role in grouping big data in various implementations of big data grouping. Fuzzy clustering method provides more flexibility than non-fuzzy methods by allowing each data record to belong to more than one cluster to a certain level [14].

Among the fuzzy grouping methods, fuzzy c-means (FCM) is the most famous method. In this method algorithm, it is assumed that all features are equally important.
In this paper, we present the existence of fuzzy c-means algorithms using Minkowski distance and Euclidean distance, to process datasets with high dimensions and the high number of clusters [15]. Fuzzy c-means is a popular distance-based objective function. The objective function of FCM is stated by:

$$J_{m}^{FCM}(U, X, \{A_i\}) = \sum_{i=1}^{K} \sum_{k=1}^{N} (\mu_{ik})^m D_{ik}^2 A_i$$  \hspace{0.5cm} (1)$$

where

$$U = \left[\mu_{ij}\right]_{K \times N} \in [0,1]$$ \hspace{0.5cm} (2)$$

$$X = \mathbb{R}^{n \times N}$$ \hspace{0.5cm} (3)$$

with,

- $U$ = partition matrix
- $A_i$ = local norm-inducing matrix which used as one of the optimization variables
- $X$ = a set of data not labeled $p$-dimensional
- $\mu_{ij}$ = the membership degree of data object $x_k$ in cluster $K_i$, which is stated by:

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{K} (d_{ij} / d_{il})^{m-1}}; \quad i = 1,2,\ldots,K; \quad j = 1,2,\ldots,N$$ \hspace{0.5cm} (4)$$

2.2. Minkowski Distance

Minkowski distance is a distance between two points in the normed vector space and is a generalization of the Euclidean distance and Manhattan distance [16]. Suppose $X$ is a vector space then the norm on $X$ is a real-valued function $\|x\|$ which satisfies below conditions[17]:

a. Zero Vector will have zero length.

b. Scalar Factor, there will be no change in the direction of the vector when we multiply it with a positive number.

c. Triangle Inequality, the distance calculated between two points will always be a straight line.

Let $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ dan $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ to 2 points in a vector space, then the Minkowski distance between $i$ and $j$ is defined as [18]:

$$d(i, j) = \sqrt{h \left[|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{in} - x_{jn}|^h\right]}$$ \hspace{0.5cm} (5)$$

where, $h$ is a real number, such that $h \geq 1$.

2.3. Euclidean Distance

Euclidean distance is the most commonly used distance search method. In many cases, when people say about distance, they will refer to Euclidean distance [19].

Let $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ dan $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ to 2 points in a vector space, then the Minkowski distance between $i$ and $j$ is defined as [3]:

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{in} - x_{jn}|^2}$$ \hspace{0.5cm} (6)$$

2.4. Confusion Matrix

The confusion matrix is used to measure the performance of a classification algorithm. The matrix can be easily understood, although the terminology associated with mixed matrices is a bit confusing [13] (see Table 1)
Table 1. Confusion Matrix.

| Actual | Predicted | True Positive (TP) | False Positive (FP) | False Negative (FN) | True Negative (TN) |
|--------|-----------|--------------------|---------------------|---------------------|--------------------|

Note:
- **TP**: Correct Positive: Infarction is predicted, and there is an infarction
- **FP**: False Positive: Infarction is predicted, and there is no infarction
- **FN**: False Negative: There is no predictable infarction, and infarction exists
- **TN**: True Negative: There is no predictable infarction and no infarction

From the confusion matrix, we can find accuracy, accuracy, memory and score f1, with each formula being [20]:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}
\]

\[
Precision = \frac{TP}{TP + FP} \tag{8}
\]

\[
Recall = \frac{TP}{TP + FN} \tag{9}
\]

\[
f_1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}
\]

3. Experiment

The following data are data from ischemic stroke patients with cerebral infarction in their brains. Data were taken from Cipto Mangunkusumo Hospital, Indonesia can be used to build the model of Naïve Bayes Classifier (see Table 2).

Table 2. Cerebral Infarction Dataset.

| Area | Min | Max | Average | SD | Sum | Length |
|------|-----|-----|---------|----|-----|--------|
| 0.2  | -3  | 38  | 16.88   | 9.3| 5166 | 2      |
| 0.1  | 15  | 44  | 30.64   | 7.37| 7722 | 0      |
| 0.2  | -5  | 51  | 19.19   | 12.44| 4797 | 1.9    |
| ...  | ... | ... | ...     | ...| ...  | ...    |
| 0.1  | 21  | 51  | 34.67   | 6.37| 3432 | 1.6    |
| 0    | 7   | 33  | 16.69   | 8.4 | 217  | 0.7    |
| 0.1  | 25  | 59  | 44.73   | 6.16| 3847 | 1.5    |

Where there are 156 data with 7 features proportional to 70% as training data and 30% testing data from the original data, with the actual number 103 main data showing data classes with no infarction and 53 minor data indicating infarction. Table 3 shows an explanation of the infarction data features examined (see Table 3).
Table 3. The Feature of Cerebral Infarction Dataset.

| No | Feature | Definition of feature |
|----|---------|-----------------------|
| 1  | Area    | Size of area from the infarction point |
| 2  | Min     | Minimum value of infarction |
| 3  | Max     | Maximum value of infarction |
| 4  | Average | Average value of infarction |
| 5  | SD      | Standard error value of infarction |
| 6  | Sum     | Sum value of infarction point |
| 7  | Length  | Length of infarction point |

4. Result and Discussion

By using the Fuzzy C-Means with Minkowski and Euclidean Distance, we can get accuracy results with the training data as follows (See Table 4):

Table 4. Accuracy Value of Naïve Bayes Classifier.

|                  | Minkowski Distance | Euclidean Distance |
|------------------|--------------------|--------------------|
|                  | Data Training | Akurasi | Data Training | Akurasi |
| 20                | 89,02        | 20       | 82,93        |
| 30                | 89,58        | 30       | 88,89        |
| 40                | 86,18        | 40       | 86,18        |
| 50                | 64,08        | 50       | 62,14        |
| 60                | 65,85        | 60       | 76,83        |
| 70                | 90,16        | 70       | 85,25        |
| 80                | 85,37        | 80       | 92,68        |
| 90                | 40,00        | 90       | 90,00        |

From this table, we can see that the best accuracy value for Minkowski Distance is 90.16% with 70% training data, and the best accuracy value for Euclidean Distance is 92.68% with 80% training data. In this result, the data features are normalized.

By using the confusion matrix, we can find the performance of the Fuzzy C-Means Clustering with Euclidean Distance system (See Table 5).

Table 5. Confusion Matrix of Fuzzy C-Means Clustering with Euclidean Distance.

|                | Actual |          |          |
|----------------|--------|----------|----------|
|                 | Positive | Negative |
| Predicted       | 24      | 1        |
| Negative        | 2       | 14       |

From table 5 it can be seen that from 26% of actual positive samples (first column), it is estimated that 24% of them are positive with cerebral infarction in the brain and 2% negative with cerebral infarction in the brain. From 15% of actual samples negative (second column), it is estimated that 1% of them are positive with cerebral infarction in the brain and 14% negative for cerebral infarction in the brain.
From these results, we can calculate the value of accuracy, precision, recall, and f1 – score with 80% training data (See Table 6).

**Table 6.** Matrix classification of Naïve Bayes Classifier.

| Accuracy    | Precision | Recall | \( f_1 – \text{score} \) |
|-------------|-----------|--------|--------------------------|
| 92.68%      | 96%       | 63.15% | 94.12%                   |

Table 6 shows the accuracy of Fuzzy C-Means Clustering with Minkowski and Euclidean Distance in predicting whether or not someone has an infarction in his brain. It can be seen that the accuracy value is 92.68%, Precision 96%, Recall 63.15%, and F1 score 94.12%.

### 5. Conclusion

Based on experiments that have been done, it can be seen that using 80% training data produces 92.68% accuracy for Euclidean distance, and 70% produces 90.16% accuracy for Minkowski distance. To get the next results, we use Euclidean distance which has a higher accuracy, reaching 92.68%. So that obtained 96% Precision, 63.15% Recall, and F1 score 94.12%.

### Acknowledgment

This research is supported financially by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia (KEMENRISTEKDIKTI), with the PTUPT 2020 research grant scheme. And we thank the leaders of hospitals, doctors, and laboratory assistants at Cipto Mangunkusumo National General Hospital, DKI Jakarta, who have shared knowledge and experience to help this research.

### References

[1] Darotin R, Nurdiana and Nasution T H 2017 Analysis of Predictive Factors of Mortality in Hemorrhagic Stroke Patients at Soebadi Hospital Jember. *NurseLine Journal. Vol.2.No.2.*

[2] Eduardo S and Sunl W 2004 Priority Medicines for Europe and the World “A Public Health Approach to Innovation”. [Internet]. [Access on 2019]. Available at https://www.who.int/medicines/areas/priority_medicines/BP6_6Stroke.pdf

[3] Zuryati T Q and Adityo W Stroke 2016 Hemoragik e.c Hipertensi grade II Vol.5.No.2.

[4] InfoDATIN Pusat Data dan Informasi Kementrian Kesehatan RI Hipertensi. [Internet]. [Access on 2019]. Available at http://www.depkes.go.id/folder/view/01/structure-publikasi-pusdatin-info-datin.html

[5] Zuherman R, Dea A U, Jacob P and Widyo A N 2019 Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalance Cerebral Infarction Dataset Indonesia Vol.9.No.2.

[6] Brian W 2017 What’s to Know About Hemorrhagic Stroke? [Internet]. [Accesses on 2019]. Available at https://www.medicalnewstoday.com/articles/317111.php

[7] Parida H, Rahayu L and Rasmalah 2018 Hubungan Karakteristik dan Dukungan Keluarga Lansia dengan Kejadian Stroke pada Lansia Hipertensi di Rumah Sakit Umum Pusat Haji Adam Malik Medan. *Jumantik. Vol.3.No.1*

[8] The Internet Stroke Center An Independent Web Resource for Information About Stroke Care and Research. [Internet]. [Accesses on 2019]. Available at http://www.strokecenter.org/patients/about-stroke/ischemic-stroke/

[9] Negar A, Lesly A P, Makoto N, Thalia S F, Carlos B, Franco C, Leslie A M, David C A, Robert G H and Oscar G B 2014 Clinical Correlates of Infarct Shape and Volume in Lacunar Strokes. *Vol.45.No.10.*

[10] Amanda Rizki Bugasta et al 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052016
[11] Arfiani et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052011
[12] Nadisa Karina Putri et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052059
[13] Nafizatus Salmi and Zuherman Rustam 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052068
[14] Nasser G, Meysam G, Mohammad A N 2017 BigFCM: Fast, precise and scalable FCM on hadoop. Future Generation Computer Systems 77 (2017) 29-39.
[15] Natacha G, Iren V, George G 2017 M&MFCM: Fuzzy C-Means Clustering with Mahalanobis and Minkowski Distance Metrics. Procedia Computer Science 114 (2017) 224-233
[16] Minkowski Distance. [Internet]. [Access on 2019]. Available at https://iq.opengenus.org/minkowski-distance/
[17] Importance of Distance Metrics in Machine Learning Modelling. [Internet]. [Access on 2019]. Available at https://towardsdatascience.com/importance-of-distance-metrics-in-machine-learning-modelling-e51395ffe60d
[18] Jiawei H and Jian P 2012 Minkowski Distance. [Internet]. [Access on 2019]. Available at https://www.sciencedirect.com/topics/computer-science/minkowski-distance
[19] Euclidean Distance. [Internet]. [Access on 2019]. Available at https://people.revoledu.com/kardi/tutorial/Similarity/EuclideanDistance.html
[20] Durrabida Zahras et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 546 052089