Classification of Cerebral Infarction Data Using K-Means and Kernel K-Means

A M Putri1*, A G M Sari1, Z Rustam1, and J Pandelaki2

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

e-mail: ajeng.maharani@ui.ac.id

Abstract. A cerebral infarct is a circumscribed focus or area of brain tissue that dies as a result of localized hypoxia or ischemia due to cessation of blood flow. To diagnose the presence of cerebral infarction, it needs a CT scan result from the patient. But, in this study not only CT scan result will be used, machine learning also will be proposed to diagnosing cerebral infarction. Machine learning can be used to detect and classify of infarcts in the brain using features and label that obtained from the results of the CT scan. In this study, the machine learning method that will be used is K-Means and K-Means based on kernel or kernel K-Means. Kernel K-Means is the application of K-Means that modified by changing the inner product with kernel function. The CT scan result data used in this study was obtained from the Department of Radiology at Dr. Cipto Mangunkusumo Hospital (RSCM). The best result reached with kernel K-Means, it performed with different percentage of training data, started with 50%, 55%, until 95% data training. The average accuracy score of the kernel K-Means method attained an accuracy rate of 95.28%.

Keywords: Classification, cerebral infarction data, K-means, kernel K-means

1. Introduction

Stroke is the leading cause of death and disability in many countries [1]. It is a serious problem in Asia, which has more than 60% of the world’s population, and many of its countries are developing countries [2]. From Southeast Asia Medical Information Center shows that the highest mortality rate of stroke resulted from Indonesia, followed by the Philippines, Singapore, Brunei, Malaysia, and Thailand [3].

Stroke is a medical condition that the sudden death of some brain cells due to lack of oxygen when the blood flow to the brain is lost by blockage or rupture of an artery to the brain [4]. Generally, stroke can be classified into two main categories: hemorrhagic and ischemic. Hemorrhagic is a stroke caused by a weakened blood vessels that ruptures and bleed into the surrounding brain [5]. Meanwhile, ischemic stroke occurs when blood vessels that supply the brain become blocked and disrupt blood flow to parts of the brain [6]. Blocked arteries are called infarction. In ischemic stroke, cerebral infarction is the common condition and is the death of brain cells because of prolonged ischemia [7].
Cerebral infarction in ischemic stroke patient can be seen using Computed Tomography (CT) scanners. But, in this study not only CT scan result will be used, machine learning also will be proposed to diagnosing cerebral infarction. Machine learning can be used to detect and classify of infarct in the brain using features and label that obtained from the results of the CT scan. In this study, the machine learning method that will be used is K-Means and K-Means based on kernel or kernel K-Means. Kernel K-Means is the application of K-Means that modified by changing the inner product with kernel function. Kernel K-means allows recognizing more complex cluster shapes than standard K-means because of its non-linear distance function based on kernels [8].

2. Materials and Methods

2.1. Dataset

The dataset used in this paper obtained from the results of the CT scan of the Department of Radiology Cipto Mangunkusumo Hospital. Each observation consists of 7 features and a label. The label ‘1’ is used for the dataset that indicate the presence of infarction, and the label '0' is used for represent data classes that do not indicate infarction. The attributes shown in Table 1.

| Attribute | Description |
|-----------|-------------|
| Area      | Size of the area from the infarction point |
| Min       | Minimum value of infarction |
| Max       | Maximum value of infarction |
| Average   | Average value of infarction |
| SD        | Standard error value of infarction |
| Sum       | Sum value of infarction point |
| Length    | Length of infarction point |

2.2. K-Means Clustering

K-Means is an unsupervised learning algorithms that solve the clustering problem. The purpose of algorithm is to define k centroids, for each cluster. These centroids should be placed correctly because of different location can causes different result [9]. Given a set of n elements \( X = \{x_1, x_2, \ldots, x_n\} \) and c centers of the clusters, \( \{v_1, v_2, \ldots, v_c\} \). The objective function given by equation (1)

\[
J = \sum_{i=1}^{n} \sum_{j=1}^{c} A_k(x_i) \|x_i - v_j\|^2
\]



\( A_k(x_i) \in \{0,1\} \) where \( k = 1, 2, \ldots, c \) describe which c clusters the data point \( x_i \) is assigned to. The steps of K-Means algorithm are shown in Figure 1 [10] [11].
2.3 Kernel Function

The kernel function is defined as follows [12]:

\[ K(x, y) = \langle \varphi(x), \varphi(y) \rangle \tag{2} \]

Where \( \varphi(x) \) is the mapped feature space by input data \( x \). For measuring distance between \( \varphi(x) \) and \( \varphi(y) \) used the equation below [13]

\[ (d(x, y))^2 = K(x, x) - 2K(x, y) + K(y, y) \tag{3} \]

These are some common kernel functions [14]:

- Polynomial kernel
  \[ K = (x, y) = (x^T y + \gamma)^d \tag{4} \]

- Gaussian RBF kernel
  \[ K(x, y) = \exp \left( -\frac{\|x-y\|^2}{2\sigma^2} \right) \tag{5} \]

- Sigmoid kernel
  \[ K(x, y) = \tanh(y(x^T y) + \theta) \tag{6} \]

In this study, gaussian RBF kernel will be used in the next formula.

2.4 Kernel K-Means Clustering

Kernel K-Means is a K-Means algorithm that modifies the inner product into the kernel function. The objective function given by equation (7)

\[ J = \sum_{i=1}^{n} \sum_{j=1}^{c} \bar{A}_k(x_i) [K(x_i, x_i) - 2K(x_i, v_j) + K(v_j, v_j)] \tag{7} \]

\( \bar{A}_k(x_i) \in \{0, 1\} \) where \( k = 1, 2, \ldots, c \) describe which \( c \) clusters the data point \( x_i \) is assigned to.

The steps of Kernel K-Means algorithm are shown in Figure 2.
In this study, Gaussian RBF kernel where \( K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right) \) will be used.

### 3. Result and discussion

The confusion matrix is used to evaluate the performance of a model and express the classification and misclassification by the system. Accuracy score will be used to compare the performance of the methods. Formula for finding the value of accuracy shown below.

\[
\text{Accuracy} = \frac{(TN+TP)}{(FN+TP+FP+TN)} \times 100\% 
\]  

Where,

- FN: Infarction is not predicted, and infarction is present
- TN: Infarction is not predicted, and no infarction is present
- FP: Infarction is predicted, and infarction is not present
- TP: Infarction is predicted, and infarction is present

The performance of K-Means clustering and Kernel K-Means clustering was analyzed with two separate data, that is training and testing data. These methods were performed with different percentages of the data used which start from 50 to 95 percent. The results of K-Means and Kernel K-Means shown in Table 2 and Table 3.

#### Table 2. Results of Cerebral Infarction Classification Accuracy using K-Means.

| Data Training (%) | Accuracy (%) |
|-------------------|--------------|
| 50%               | 59.22        |
| 55%               | 59.14        |
| 60%               | 60.25        |
| 65%               | 57.53        |
| 70%               | 53.23        |
| 75%               | 59.62        |
| 80%               | 61.90        |
| 85%               | 61.29        |
| 90%               | 61.90        |
| 95%               | 72.73        |
Correspond to Table 2, the training data at 95% with 72.73% accuracy was recorded as the best accuracy. Meanwhile, the lowest accuracy result was recorded at data training 70% with 53.23% accuracy. The result when we use the Kernel K-Means clustering shown in Table 3.

**Table 3.** Results of Cerebral Infarction Classification Accuracy using Kernel K-Means with Gaussian RBF Kernel, parameter: $\sigma = 20$.

| Data Training (%) | Accuracy (%) |
|-------------------|--------------|
| 50%               | 98.06        |
| 55%               | 97.34        |
| 60%               | 98.38        |
| 65%               | 97.76        |
| 70%               | 96.50        |
| 75%               | 92.90        |
| 80%               | 93.29        |
| 85%               | 94.28        |
| 90%               | 92.43        |
| 95%               | 91.84        |

**Table 4.** The Results of The Accuracy and The Running Time of Prostate Cancer Classification by Support Vector Machines.

| Data Training | Accuracy | Running Time |
|---------------|----------|--------------|
| 0.1           | 81.92%   | 0.06s        |
| 0.2           | 66.89%   | 0.06s        |
| 0.3           | 66.67%   | 0.04s        |
| 0.4           | 66.67%   | 0.05s        |
| 0.5           | 63.04%   | 0.04s        |
| 0.6           | 66.22%   | 0.06s        |
| 0.7           | 67.27%   | 0.05s        |
| 0.8           | 72.97%   | 0.06s        |
| **0.9**       | **83.33%**| **0.05s**    |

From the result in Table 3, we can see that Kernel K-Means has better results with Gaussian RBF using $\sigma = 20$. The best accuracy recorded when the data training at 60% with 98.38% accuracy.

**4. Conclusions**

Predicting the presence of cerebral infarction for patients can assist radiologist to diagnose the ischemic stroke. In this study, K-Means clustering and kernel K-Means clustering has been used to approach in diagnosing the presence of cerebral infarction.

The comparison between K-Means and kernel K-Means can be seen that K-Means based on kernel has better results. K-Means without kernel has average accuracy 60.68% with 50%-95% data training, and the best accuracy recorded at 95% data training with 72.73% accuracy. Meanwhile, kernel K-Means has average accuracy 95.28% with 50%-95% data training and $\sigma = 20$, and the best accuracy recorded at 60% data training with 98.38% accuracy.
For future research, it is possible to use another kernel function to see what kind of kernel functions work well with the data, or use another method to increase the accuracy so the classification can be used for better diagnosis.

5. Acknowledgment
This research supported financially by the Ministry of Research, Technology, and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with PTUPT 2020 research grant scheme. We would like to thank Dr. dr. Jacub Pandelaki, Sp.Rad(K) from Department of Radiology, Cipto Mangunkusumo Hospital for the dataset.

References
[1] Narayanaswamy V, Byung WY, Jeyaraj P, Jose CN 2017 Stroke Epidemiology in South, East, and South-East Asia: A Review J Stroke. 2017 Sep; 19(3): 286–294.
[2] Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, et al 2010 Global Burden of Diseases and Risk Factors Study (GBD 2010) and the GBD Stroke Experts Group Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet. 2014;383:245–254.
[3] Zuherman R, Dea AU, Rahmat H, Jacob P, Widyo AN 2019 Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infarction Datasets. Vol.9 (2019) No. 2 ISSN: 2088-5334
[4] Walter J, Oyere O, Mayowa O, Sonal S 2016 Stroke: a global response is needed. Bull World Health Organ 2016;94:634–634A. doi: http://dx.doi.org/10.2471/BLT.16.181636
[5] American Stroke Association: Hemorrhagic Stroke. https://www.stroke.org/en/about-stroke/types-of-stroke/hemorrhagic-strokes-bleeds. Accessed on 4 October 2019
[6] John Hopkins Medicine: Types of Stroke. https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/types-of-stroke. Accessed on 4 October 2019
[7] I.A.Mentari, R.Naufalina, M.Rahmadi, J.Khotib 2018 Development of Ischemic Stroke Model By Right Unilateral Common Carotid Artery Occlusion (RUCCAO) Method Fol Med Indones, vol.54, no.3, pp.200-206, 2018.
[8] Sebastian M, Emilio C, Richard W 2015 Kernel Penalized K-means: A feature selection method based on Kernel K-means. Information Sciences Vol. 322, 20 November 2015, Pages 150-160
[9] Velmurugan T 2014 Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data. Applied soft computing. p 134-146
[10] Israa AA, Adel M, S Mahmoud T 2018 K C – Means: A Fast Fuzzy Clustering. Advances in Fuzzy Systems; New York. DOI:10.1155/2018/2634861. p 137
[11] Zuherman R and Sri H 2019 Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel. IOP Conf. Series: Materials Science and Engineering. doi:10.1088/1757899X/546/5/052067.
[12] Arfiani, Zuherman R, Jacob P, Arga S 2019 Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification. IOP Conf. Series: Materials Science and Engineering546 (2019) 052011. doi:10.1088/1757-899X/546/5/052011.
[13] Christopher B 2006 Pattern Recognition and Machine Learning. New York: Springer
[14] Grigorios T and Aristidis L 2008 The Global Kernel k-Means Clustering Algorithm. Conference: Neural Networks, 2008. IJCNN 2008. DOI: 10.1109/IJCNN.2008.4634069.