Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

The Global Kernel *k*-means Clustering Algorithm for Cerebral Infarction Classification

To cite this article: Z Rustam *et al* 2019 *J. Phys.: Conf. Ser.* **1417** 012027

View the article online for updates and enhancements.

You may also like

- Naïve bayes classifier models for cerebral infarction classification
SG Fitri, R Selsi, Z Rustam et al.
- On the feasibility of using the spiral beam formalism for analysis of cardiograms
V.G. Volostnikov, S.A. Kishkin, S.P. Kotova et al.
- Detecting and interpreting myocardial infarction using fully convolutional neural networks
Nils Strodthoff and Claas Strodthoff

This content was downloaded from IP address 79.25.203.58 on 26/11/2022 at 20:16

The Global Kernel k-means Clustering Algorithm for Cerebral Infarction Classification

Z Rustam[1*], S G Fitri [1], R Selsi[1], J Pandelaki[2]

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

e-mail: *rustam@ui.ac.id

Abstract. Cerebral infarction is the death of neurons, glia cells and blood vessel systems caused by a lack of oxygen and nutrients. This situation is often called storke. Common causes of neuron damage are hypoxia, which is caused by impaired blood flow, reduced oxygen pressure in blood circulation, toxins, and hypoglycemia which can result in the same morphological changes as morphological changes in hypoxia. Hypoxia is reduced oxygen pressure in the alveoli, resulting in hypoxemia which can cause hypoxic brain tissue. The initial stage of ischemic neurons is characterized by the formation of micro vacuolization, which is characterized by the size of the cells that are still normal or slightly reduced, the nucleus shrinks slightly, vacuoles occur in the perikaryon region. This micro vacuole can be found in neurons in hippokamus and cortical 5-15 minutes after hypoxia. The final sign of cell damage due to ischemia is characterized by the nucleus becoming pyknotic and fragmented. To classify cerebral infarction, the author uses the global k-means clustering algorithm as a classification method that shows that the method has good accuracy, good memory, and good precision in classifying cerebral infarction. In this proposed method, the global kernel k-means clustering algorithm is an extension of the standard k-means clustering algorithm and has been used to identify or classify clusters that are non-linearly separated in space input. This method adds one cluster at each stage through a global search process consisting of several k-means kernel executions from the appropriate initialization. Therefore, this method can make good classification accuracy. In particular, this achieves classification accuracy of up to 78% for the highest accuracy.

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 caused by blockages (infarction) or rupture of blood vessels that carry oxygen and nutrients to the brain [2] so that the brain does not get the blood and oxygen needed and causes the death of brain cells [3]. 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 [4].

Hypertension is the most important risk factor [5], which can increase the risk of stroke by about two to four times [6]. Increased blood pressure will cause cerebral constriction. If blood pressure rises

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

in a period of months or longer than years, it will be accounted for in the cerebral blood muscle layer so that the blood vessel's lumen diameter will remain. If that happens, it can cause the cerebral arteries to not constrict to overcome fluctuations in systemic blood pressure [6].

Stroke is classified into two types, namely hemorrhagic stroke and non-hemorrhagic (ischemic) stroke [7]. 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 [8]. Whereas ischemic stroke is a stroke caused by thrombotic or thromboembolic blockages in the arteries [5]. This blockage of the arteries to the brain causes disruption of blood flow to the brain so that brain cells cannot make enough energy and will stop working [9].

Ischemic stroke is not only caused by clogged arteries to the brain, but can also be caused by the use of drugs, blood clotting disorders, or traumatic damage to neck veins [9]. The number of stroke patients is estimated to reach 85% of the number of strokes that occur [10]. Ischemic stroke is a major cause of disability in the world. In ischemic stroke, cerebral infarction is a more common condition [7]. Cerebral infarction is the death of neurons, blood cells and blood vessel systems caused by a lack of oxygen and nutrients [11].

Stroke affects each person differently, there are some people who need more time to recover [9]. Recovery from physical, social and emotional changes. Treatment for stroke usually starts with hospital treatment. Stroke rehabilitation requires good coordination between patients, families, doctors, nurses, physical therapists, psychologists, and others. Aspects of certain aspects of stroke rehabilitation have been well established in clinical practice and are a standard of care, for example providing physical therapy for patients with early stroke [9].

Treatment or 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. In addition, strokes need to be avoided, also need to modify the lifestyle of not smoking, not drinking alcohol and not using cocaine [6]. There are several studies that have addressed this resampling technique, including Z Rustam et al., Which classifies unbalanced datasets to predict cerebral infarction. The modeling technique used is Hybrid Preprocessing Method for Support Vector Machine[7].

In this study, we used the Global Kernel k-means Clustering Algorithm to classify datasets of cerebral infarction in the brain using data obtained from the Radiology Laboratory at Cipto Mangunkusumo Hospital, Indonesia. In this proposed method, the global k-means clustering algorithm is an extension of the standard k-means clustering algorithm and has been used to identify or classify clusters that are non-linearly separated in space input. This method adds one cluster at each stage through a global search process consisting of several k-means kernel executions from the appropriate initialization. By using the k-means kernel, the algorithm works in stages by solving all problems $1, \ldots, M$ clusters. This algorithm combines the global advantages of k-means and the k-means kernel so as to avoid the limitations of both k-means [12].

2. Method

2.1 Clustering

Clustering aims to group similar patterns such as features or data items into certain categories and specifically for classification. Many grouping algorithms have been designed to meet needs, such as efficiency and accuracy, each of which has advantages and disadvantages[13].

2.2 Kernel K-Means Clustering

To group data that is not liner inseparable in the original feature space, the k-means grouping is extended to the kernel version. However, the workings of the k-means kernel depend on the choice of kernel functions, so many kernel learning has been introduced to k-means clustering to get the optimal kernel combination[14].

Some basic structures are taken from the existing k-means grouping approach. The resulting kernel k-means clustering algorithm is[13]:

- Initialization: K is chosen randomly in the dataset as the initial centroid c_p, $1 \leq p \leq K$

- Clustering: Similarity $Sim\ (r_i, c_p)$ between each record and centroid is calculated, and record r_i record is assigned to the c_p centroid which is similar.
- Centroid re-calculation: c_p the new centroid is recalculated for each cluster with the average embedding vector of all r_i records assigned to the cluster, namely:
-

$$c'_p = \frac{1}{nc_p} \sum_{i=1}^{nc_p} \quad r_i \tag{1}$$

Where, nc_p is the number of records in the cluster indicated by the centroid c_p

- Termination conditions: the matrix for all old centroids is denoted by $\Theta = \{c_1, \dots, c_p, \dots, c_k\}^T, 1 \le p \le K$ and the matrix for all new centroids is denoted by $\Theta' = \{c'_1, \dots, c'_p, \dots, c'_k\}^T$. Next, Euclidean distance $d(\Theta, \Theta')$:

$$d(\Theta, \Theta') = \sqrt[2]{\sum_{p=1}^{k} \quad (c_p - c'_p)^2} = \sqrt[2]{\sum_{p=1}^{k} \quad \sum_{i=1}^{\lambda} \quad (v_{p,i} - v'_{p,i})^2} \tag{2}$$

Where $v_{p,i}$and $v'_{p,i}$ with i are daro centroid elements c_p and c'_p.

2.3 K-Means and Global K-Means

K-means is one of the most popular grouping algorithms.

Suppose we are given $X = \{x_1, x_2, \dots, x_N\}, x_N \in R^d$, The dataset will be divided into separate M, namely $C_1, C_2, \dots, C_M$. The algorithm finds the optimal local solution with respect to grouping errors as the sum of the square of the Euclidean distance between each data. Analytically the grouping errors are stated by[15]:

$$E(m_1, \dots, m_M) = \sum_{i=1}^{N} \quad \sum_{k=1}^{M} \quad I(x_i \in C_k) \|x_i - m_K\|^2 \tag{3}$$

The disadvantages of the k-means algorithm are:

- The final solution depends on the starting position of the cluster center
- Clusters are separated linearly

To overcome this problem, we use the global k-means algorithm[15]. Global k-means is a data grouping approach that dynamically adds one cluster center at once by using a deterministic global from the initial position consisting of N data sizes from the k-means algorithm[16].

2.4 Kernel K-Means

The kernel k-means is a generalization of the k-means standard algorithm, where data points are mapped from space input to higher dimensional feature space through non-linear transformation ∅. In the k-means kernel, data is expected to separate well because overlapping data can be linear in new dimensional space [15] (see table 1).

Table 1. Examples of kernel functions

Polynomial kernel	$K(x_i, x_j) = (x_i^T x_j + \gamma)^\delta$
Gaussian kernel	$K(x_i, x_j) = exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)$
Sigmoid kernel	$K(x_i, x_j) = tanh(\gamma(x_i^T x_j) + \theta)$

The objective function to be reduced by the kernel k-means is equivalent to the grouping error in the feature, which is expressed by[15]:

$$E(m_1, \dots, m_M) = \sum_{i=1}^{N} \quad \sum_{k=1}^{M} \quad I(x_i \in C_k)\|\emptyset(x_i) - m_k\|^2 \tag{4}$$

Where, $m_k = \frac{\sum_{i=1}^{N} \; I(x_j \in C_k)\emptyset(x_i)}{\sum_{i=1}^{N} \; I(x_j \in C_k)}$

$$\|\emptyset(x_i) - m_k\|^2 = K_{ii} - \frac{2\sum_{j=1}^{N} \; I(x_i \in C_k)K_{ij}}{\sum_{j=1}^{N} \; I(x_i \in C_k)} + \frac{\sum_{j=1}^{N} \; \sum_{l=1}^{N} \; I(x_j \in C_k)I(x_l \in C_k)K_{jl}}{\sum_{j=1}^{N} \; \sum_{l=1}^{N} \; I(x_j \in C_k)I(x_l \in C_k)} \tag{5}$$

In this study, we use the Radial Basis Function Kernel (RBF Kernel). Conventional RBF networks employ a number of kernels such as multi-quadrics, inverse multi-quadrics, and Gaussian, with the following formula [18]:

$$K(x_i, x_j) = exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{6}$$

2.5 Global Kernel K-Means

The global kernel k-means is a combination of global k-means and kernel k-means.
The global kernel k-means minimizes grouping errors in the feature space. The global kernel k-means map datasets from the input space to the higher dimensional feature space using the kernel matrix. Global kernel k-means produces optimal solutions in clustering problems[17].

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 The Global Kernel K-Means Clustering (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,	18	51	32.29	7.84	8136	1.8
0.1	-14	26	8.67	10.25	824	1.2
0.1	25	61	38.99	7.37	6122	1.5
...	...	...	...	...	...	...
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 4 shows an explanation of the infarction data features examined.

Table 3. The Feature of Cerebral Infarction Dataset

No	Feature	Definition of feature
1	Area	The size of the area from the infarction point
2	Min	The minimum value of infarction
3	Max	The maximum value of infarction
4	Average	The average value of infarction
5	SD	Standard error value of infarction
6	Sum	The sum value of infarction point
7	Length	Length of infarction point

4. Result and Discussion

By using the Global Kernel K-Means Clustering Algorithm, it can be obtained as follows, where k-fold divides the sample data into k parts. To get the result, we use a Radial Basis Function Kernel (RBF Kernel).

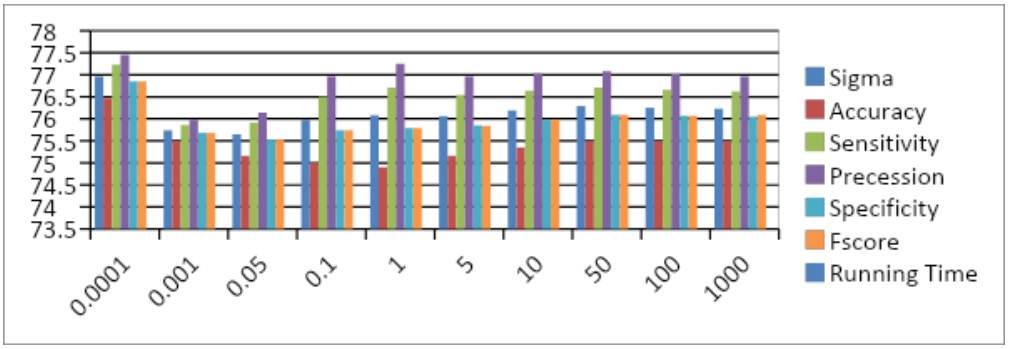

Figure 1. For k-fold= 3

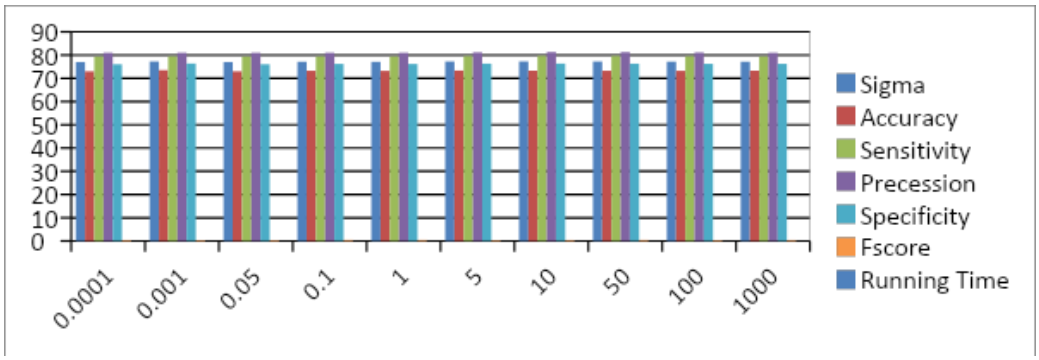

Figure 2. For k-fold= 5

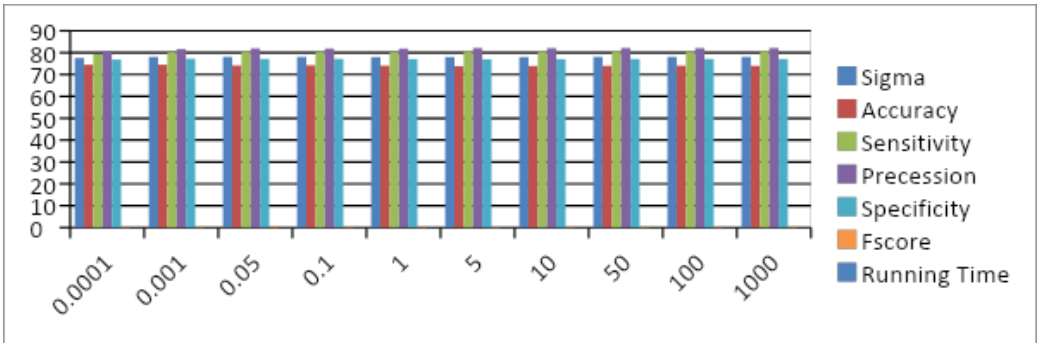

Figure 3. For k-fold= 7

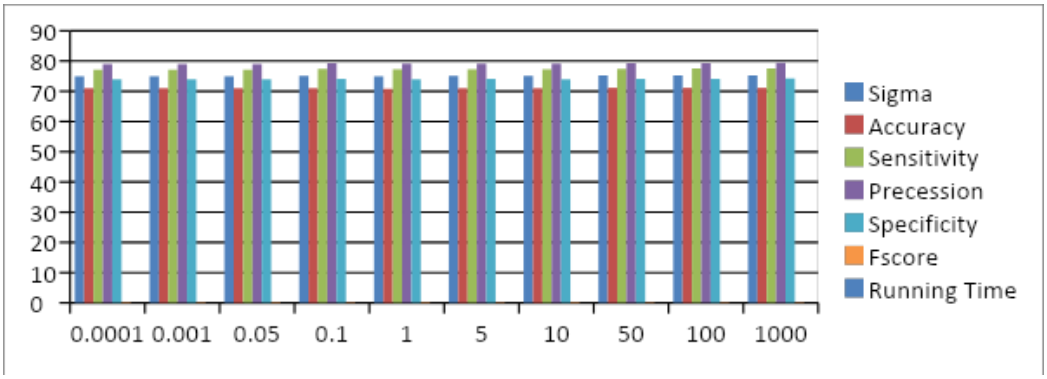

Figure 4. For k-fold= 10

From the chart, we can see that for k-fold=3, the best accuracy is 76.96% while the sigma is 0.0001. For k-fold= 5, the best accuracy is 77.28% while the sigma is 10. For k-fold= 7, the best accuracy is 78.06% while the sigma is 0.001. And last, for k-fold= 10, the best accuracy is 75.25% while the sigma is 1000. So, for all of the k-fold, k-fold=7 with sigma 0.001 gives the best accuracy value, which is 78.06%. While the sensitivity value is 74%, the precession value is 80%, specificity value is 81% and FScore 77%.

5. Conclusion

From the experiments conducted in section 4, we can see that the accuracy achieved by the K-Means Clustering Algorithm Global Kernel method is 78.06%. This method provides high accuracy to classify whether there is an infarction in a person's brain or not. The disadvantage of this method is that it does not produce high accuracy values in terms of classification so that further research is needed by using new methods or with other kernel functions so that it will produce better accuracy.

Acknowledgment

This research is supported financially by the University of Indonesia, with DRPM PIT-9 2019 research grant scheme, ID number NKB-0039/UN2.R3.1/HKP.05.00/2018

References

[1] Darotin R, Nurdiana, and Nasution TH. 2017 Analysis of Predictive Factors of Mortality in Hemorrhagic Stroke Patients at Soebadi Hospital Jember *NurseLine Journal* **2** (2)

[2] Bay V, Kjolby B F, and Iversen N K 2018 Stroke Infarct Volume Estimation in Fixed Tissue: Comparison of Diffusion Kurtosis Imaging to Diffusion-Weighted Imaging and Histology in a Rodent MCAO Model *PLoS ONE* **13** (4) e0196161

[3] Wang G, Jing J, and Pan Y 2019 Does All Single Infarction have Lower Risk of Stroke Recurrence Than Multiple Infarctions in Minor Stroke? *BMC Neurology* **19** (7)

[4] Darotin R, Nurdiana, and TH Nasution 2017 Analysis of Predictive Factors of Mortality in Hemorrhagic Stroke Patients at Soebadi Hospital Jember *NurseLine Journal* **2** (2)
[5] Hanum P, Lubis R, and Rasmalian - *Hubungan Karakteristik dan Dukungan Keluarga Lansia dengan Kejadian Stroke pada Lansia Hipertensi di Rumah Sakit Umum Pusat Haji Adam Malik Medan* (Medan: RSU Pusat Haji Adam Malik)
[6] Qurbany Z T and Wibowo A - *Stroke Hemoragik e.c Hipertensi Grade* II Indonesia
[7] Rustam Z, Utami D A, Pandelaki J, and Nugroho WA - *Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalance Cerebral Infarction Dataset* Indonesia
[8] Wu B 2017 *What's to Know About Hemorrhagic Stroke?* Available at https://www.medicalnewstoday.com/articles/317111.php
[9] The Internet Stroke Center 2019 *An Independent Web Resource for Information About Stroke Care and Research* Available at http://www.strokecenter.org/patients/about-stroke/ischemic-stroke/
[10] Hadayani D and Dominica D 2018 Gambaran Drug Related Problems (DRP'S) pada Penatalaksanaan Pasien Stroke Hemoragik dan Stroke Non-Hemoragik di RSUD DR M Yunus Bengkulu. *Jurnal Farmasi dan Ilmu Kefarmasian Indonesia* **5** (1)
[11] Japardi I 2002 *Neuropatologi Infark Serebi* Indonesia
[12] Tzortzis G F and Aristidis C 2019 The Global Kerel K-means Algorithm for Clustering in Feature Space *IEEE Transactions on Neural Networks* **20** (7)
[13] Zhang Y, Lu J, Liu F, Liu Q, Porter A, Chen H, Zhang G 2018 Does Deep Learning Help Topic Extraction? A Kernel K-Means Clustering Method With Word Embedding *Journal of Informetrics*
[14] Yao Y, Chen H 2018 *Multiple Kernel K-Means Clustering bt Selecting Representative Kernels* Available at https://arxiv.org/abs/1811.00264
[15] Tzortzis G and Likas A 2008 *The Global Kernel K-Means Clustering Algorithm* Greece
[16] Agrawal A and Gupta H 2013 Global K-Means (GKM) Clustering Algorithm: A Survey *International Journal of Computer Applications* **79** (2)
[17] Tzortzis G F and Likas A C 2009 The Global Kernel K-Means Algorithm for Clustering *Feature Space* **20** (7)
[18] Khan S, Naseem I, Togneri R, and Bennamoun M 2019 A Novel Adaptive Kernel for the RBF *Neural Networks*