Breast cancer clustering using modified spherical K-Means

Zuherman Rustam¹ and Ajeng Leudityara Fijri²
¹,²Department of Mathematics, Universitas Indonesia Depok, 16424, Indonesia

1 Rustam@ui.ac.id, 2Ajeng.leudityara@sci.ui.ac.id

Abstract. Clustering is one of common techniques to group dataset into subsets based on distance measure. It has been applied in machine learning, pattern recognition, data mining, image analysis, and bioinformatics. Spherical k-means is one of clustering methods to address computational efficiency and solution quality in terms of deciding an action. In this paper, we used modified spherical k-means by using kernel radial basis function (RBF) by inner product measures in spherical k-means to cluster breast cancer Coimbra dataset from UCI machine learning into clusters. A new clusters will defined to healthy control cluster and patient cluster based on medical records. The highest accuracy results of kernel spherical k-means (SPKM) clustering method with radial basis function (RBF) kernel in breast cancer Coimbra (BCC) dataset is 72.41%. Addition of kernel to spherical k-means makes the results of accuracy be stable than using spherical k-means.

1. Introduction

Clustering is unsupervised learning for grouping dataset into subsets based on dissimilarity type of data. These dissimilarities can be computes to predicted issues such as predicted cancer. It will helped As we know that cancer is the disease caused by uncontrolled cells or cells growth out of control in the part of the body. The leading causes of death cancer among women in the world is breast cancer. Breast cancer is the abnormalities of cells growth in breast. The abnormal cells in breast form as thickening of breast or lump and changes of skin breast or nipple. The cells of cancer in breast can spread to other areas of the body and causes of death. World Helath Organization (WHO) estimated 9.6 million death because of cancer in 2018, there are 2.09 million women breast cancer cases and it will increase every year. According to current evidence, 30% until 50% of cancer deaths could be prevented by modifying or avoiding key risk factors[1]. Another way to prevent the cancer deaths is screening in breast or part of body with cancer cells and checkup routine. Machine learning helps the medical staff to ensure greater treatments to reduce cancer cells in the body [2]. It has been applied for another types of cancer [3-4].

K-means is one of simplest clustering techniques which commonly used for data analysis. Its partitioning the dataset into K cluster based on euclidean distance [5]. K-means is unit vector and can be considered as points in hypersphere in high dimensional. Because of the points located in the hypersphere, then this algorithm can be called as spherical k-means [6]. It is based on cosine similarity, which performs better than k-means algorithm for high dimensional data [7]. In this paper, we used improvement of spherical k-means algorithm by adding radial basis function (RBF) kernel and spherical k-means algorithm for comparison algorithm. Addition of kernel function in spherical k-means aims to get higher accuracy and efficiency. In this paper, we use RBF kernel with \( \sigma = 0.001 \) as the parameter and breast cancer Coimbra dataset from UCI machine learning repository. The breast...
The rest of this paper is organized as follows. In the section 2, we explain the recent methodology spherical k-means and kernel spherical k-means. Section 3 describes dataset and presents of experiment results, and the last but not least section 4 is conclusion of this paper.

2. Kernel Spherical K-Means

K-means is one of unsupervised learning to partitioning \( x = (x_1, x_2, x_3, \ldots, x_n) \) data in euclidean space \( \mathbb{R}^d \) into \( k \) (1 < \( k < n \)) cluster by minimizing the mean squared error of the objective standard k-means function [9-10]

\[
E = \frac{1}{N} \sum_x \| x - \mu_{k(x)} \|^2
\]  
(1)

Where \( k(x) = \arg \min_k \| x - \mu_k(x) \| \) is index of center of kth cluster to \( x \), \( N \) is the number of total data vectors, and \( \mu_k(x) \) is the unit length of center vectors.

Spherical K-means (SPKM) algorithm is the k-means in hypersphere to classify high dimensional data. It introduced by Shi Zong in 2005. SPKM is k-means algorithm using cosine similarity which measures the equation between vector data through inner product [6]. Its maximized the cosine similarity objective [6].

\[
L = \sum_x x^T V_k
\]  
(2)

Where

\[
k_n(x) = \arg \max x^T V_k
\]  
(3)

\( k_n(x) \) is the nearest centroid cluster for each vector

\[
V_k = \frac{\sum_{x \in k \cdot x} x}{{\| \sum_{x \in k \cdot x} x \|}}
\]  
(4)

The procedure of spherical k-means (SPKM) shown in algorithm 1.

2.1 Algorithm 1

Spherical K-means

**Input:** Dataset \( X = (x_1, x_2, x_3, \ldots, x_n) \) in \( \mathbb{R}^d \) and number of clusters \( K \)

**Step 1.** Initialize the unit length cluster centroid vectors \( \mu_{k(x)} \)

**Step 2.** Compute \( k_n(x) \) for each data vector \( x_n \) according by (3);

**Step 3.** Compute \( V_k \) the centroid cluster according by (4);

**Step 4.** If \( k_n(x) = k_{n+1}(x) \) then stop, otherwise go to step 2.

**Output:** Partition of the data vectors.

Kernel is the function to connect nonlinear data with machine learning algorithm. Its helped to get high accuracy in machine learning while the data is nonlinearly separable[11]. Kernel function is defined by [12].
\[ K(x_i, v_k) = \langle \emptyset(x_i), \emptyset(v_k) \rangle \]  

(5)

The kernel distance defined by [12].

\[ d^2(x_i, v_k) = K(x_i, x_i) - 2(x_i, v_k) + K(v_k, v_k) \]  

(6)

There are types of kernel function such as linear kernel, polynomial kernel, and Radial Basis Function (RBF) kernel. In this paper, we used RBF kernel because RBF is the common kernel that usually used. The RBF kernel is defined by [12]

\[ K(x_i, v_k) = \exp\left(\frac{-\|x-x_k\|^2}{\sigma^2}\right) \]  

(7)

with \(\sigma \neq 0\).

Accordingly of kernel definition, we apply kernel to spherical k-means to get high accuracy and efficiency in spherical k-means method. Kernel spherical k-means is expansion of spherical k-means by maximize equation (2) with change inner product in equation (3) to RBF kernel.

\[ k_n(x) = \arg\max_k K(x_i, v_k) \]  

(8)

\[ k_n(x) = \arg\max_k \exp\left(\frac{-\|x-x_k\|^2}{\sigma^2}\right) \]  

(9)

The procedure of kernel spherical k-means (KSPKM) shown in Algorithm 2

2.2 Algorithm 2

Kernel Spherical K-means

**Input:** Dataset \( X = (x_1, x_2, x_3, \ldots, x_n) \) in \( \mathbb{R}^d \) and number of clusters \( K \)

**Step 1.** Initialize the unit length cluster centroid vectors \( \mu_k(x) \)

**Step 2.** Compute \( k_n(x) \) for each data vector \( x_n \) according by (9);

**Step 3.** Compute \( V_k \) the centroid cluster according by (4);

**Step 4.** If \( k_n(x) = k_{n+1}(x) \) then stop, otherwise go to step 2.

**Output:** Partition of the data vectors.

3. Experiment Result

In this section, about the breast cancer Coimbra dataset and the results of spherical k-means and kernel spherical k-means clustering method.

3.1. Dataset.

The dataset used in this paper is breast cancer Coimbra which is collected from UCI machine learning repository. M. Patricio et al. collect from faculty of medicine of the university of Coimbra and hospital center of Coimbra in 2018 [8]. Breast cancer Coimbra dataset have 116 instances and 10 attributes
include the label. The label of BCC are healty controls and patients. The attributes in Breast Cancer Coimbra shown in table 1.

**Table 1. Attributes in Breast Cancer Coimbra**

| No. | Name Attributes | Units |
|-----|-----------------|-------|
| 1.  | Ages            | years |
| 2.  | Body Mass Index /BMI | kg/m² |
| 3.  | Adiponectin     | μg/mL |
| 4.  | Glucose         | mg/dL |
| 5.  | Insulin         | μU/mL |
| 6.  | HOMA            |       |
| 7.  | Leptin          | ng/mL |
| 8.  | Resistin        | ng/mL |
| 9.  | MCP-1           | pg/dL |

3.2. Results.

Table 2 shown the results using spherical k-means and kernel spherical k-means. As seen in the table II, the experiments done using 10%-90% data training of breast cancer Coimbra dataset. Highest accuracy results using Spherical K-Means is 81.82% in 80% data training with 0,16 seconds running time. The faster running time, 0,06 seconds in 30% data training and get accuracy 61.25%. The experiments done using Kernel Spherical K-Means with sigma (σ) 0.001 as the parameter of the radial basis function (RBF) kernel. The highest accuracy of kernel spherical k-means is 72.41% which are in 10% , 20% , 30% , and 60% data training with 1,19 seconds, 0,89 seconds, 0,91 seconds, and 0,89 seconds running time. The fastest running time using Kernel Spherical K-Means is 0,89 seconds in 20% and 60% data training and produce highest accuracy 72,41%.

**Table 2. Results obtained using Spherical K-Means and Kernel Spherical K-Means with σ = 0.001 of the RBF kernel**

| Data Training (%) | Spherical K-Means | Kernel Spherical K-Means |
|-------------------|-------------------|--------------------------|
|                   | Accuracy (%)      | Running Time (s)         | Accuracy (%) | Running Time (s) |
| 10                | 62,14             | 0,14                     | 72,41        | 1,19              |
| 20                | 60,87             | 0,13                     | 72,41        | 0,89              |
| 30                | 61,25             | 0,06                     | 72,41        | 0,91              |
| 40                | 65,22             | 0,09                     | 68,97        | 0,98              |
| 50                | 68,97             | 0,09                     | 70,69        | 0,94              |
| 60                | 42,22             | 0,13                     | 72,41        | 0,89              |
| 70                | 50,00             | 0,16                     | 70,69        | 0,92              |
| **80**            | **81,82**         | **0,16**                 | 68,97        | 1,06              |
| 90                | 72,73             | 0,20                     | 67,24        | 0,95              |

Table 3 shown confusion matrix of spherical k-means and table 4 shown confusion matrix of kernel spherical k-means in the highest accuracy of each method. In table 4, there are 52 healthy controls
classified into healthy controls cluster or it be called as TP (True Positive), 42 patients classified into patients cluster or it be called as TP (True Negative), and 22 patients classified into healthy controls cluster or it be called as FP (False Positive). In table 4, there are 32 healthy controls classified into healthy controls cluster or it be called as TP (True Positive), 52 patients classified into patients cluster or it be called as TN (True Negative), 20 healthy controls classified into patients cluster or it be called as TN (True Negative) and 12 patients classified into healthy controls cluster or it be called as FP (False Positive). So, the clustering accuracy calculated by[13]

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

(10)

Table 3. Confusion matrix of Spherical K-Means (SPKM)

| Actual     | Predicted  |
|------------|------------|
| Healthy controls | 52         | 0          |
| Patient    | 22         | 42         |

Table 4. Confusion matrix of Kernel Spherical K-Means (KSPKM)

| Actual     | Predicted  |
|------------|------------|
| Healthy controls | 32         | 20         |
| Patient    | 12         | 52         |

4. Conclusion
This paper proposed to clustering breast cancer dataset using modified spherical k-means by adding kernel to innerproducat measure in spherical k-means method. We use radial basis function (RBF) kernel with sigma 0.001.

Based on the results in table II, spherical k-means given higher accuracy than kernel spherical k-means and the running time of spherical k-means faster than kernel spherical k-means. But, spherical k-means reach higher accuracy in high number of data training while kernel spherical k-means get high accuracy in few number of data training. As seen on table II, kernel spherical k-means provides stable accuracy values of each data training compared spherical k-means accuracy values. In the conclusion, kernel spherical k-means is good classifier for classifying breast cancer Coimbra dataset. This method can help medical staff classify to predict breast cancer data easily.

For further work, this research can be continue using other sigma (\(\sigma\)) value or type kernel to solve breast cancer Coimbra clustering problems.

5. Acknowledgement
This research supported financially by University of Indonesia, with a DRPM PITTA 2018 research grant scheme.
References

[1] C.M.T. Louise, and N.Jodrell 1997 Screening and breast cancer: the role of breast awareness. Journal of Cancer Nursing 1(2) pp 76-80.
[2] World Health Organization. http://www.who.int/cancer/en/. access in September 19th 2018.
[3] Rachman, A.A., Rustam, Z. Cancer classification using Fuzzy C-Means with feature selection. Proceedings - 2016 12th International Conference on Mathematics, Statistics, and Their Applications, ICMSA 2016, AAC Dayan Dawood, Syiah Kuala University Banda Aceh; IndonesiaDOI: 10.1109/ICMSA.2016.7954302
[4] Panca, V., Rustam, Z. 2017 Application of machine learning on brain cancer multiclass classification AIP Conference Proceedings, Volume 1862 DOI: 10.1063/1.4991237
[5] Tom Mitchell 1997 Machine Learning McGraw Hill
[6] Zhong, Shi. 2005 Efficient Online Spherical K-means Clustering Canada:Montreal
[7] C.C. Aggarwal , A. Hinneburg , D.A. Keim 2001 On the surprising behavior of distance metrics in high dimensional space, Lect. Notes Comput. Sci.
[8] M. Patricio, J. Pereira, J. Crisostomo, P. Matafome, M. Gomes, R. Seica, and F. Caramelo, Using Resistin, glucose, age, and BMI to predict the presence of breast cancer, 2018 Jan 4;18(1):29. doi: 10.1186/s12885-017-3877-1.
[9] E. Forgy, “Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications,” Biometrics, vol. 21, no. 3, pp. 768, 1965.
[10] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Math. Statistics and Probability, 1967, pp. 281–297
[11] Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press.
[12] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. New York: Springer.
[13] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2001) 107-145