Sleep stages identification in patients with sleep disorder using k-means clustering

M U Fadhlullah*, A Resahya, D F Nugraha and I N Yulita
Universitas Padjajaran, Jatinangor, Indonesia

*Corresponding author’s e-mail: Ufadhlullah@gmail.com

Abstract. Data mining is a computational intelligence discipline where a large dataset processed using a certain method to look for patterns within the large dataset. This pattern then used for real time application or to develop some certain knowledge. This is a valuable tool to solve a complex problem, discover new knowledge, data analysis and decision making. To be able to get the pattern that lies inside the large dataset, clustering method is used to get the pattern. Clustering is basically grouping data that looks similar so a certain pattern can be seen in the large data set. Clustering itself has several algorithms to group the data into the corresponding cluster. This research used data from patients who suffer sleep disorders and aims to help people in the medical world to reduce the time required to classify the sleep stages from a patient who suffers from sleep disorders. This study used K-Means algorithm and silhouette evaluation to find out that 3 clusters are the optimal cluster for this dataset which means can be divided to 3 sleep stages.

1. Introduction
Sleep apnoea is a serious sleep disorder that occurs when a person’s breathing is interrupted during sleep. People with sleep apnoea can stop breathing during their sleep which mean they may not get enough oxygen for the body. Peretz Lavie, Paula Herer, and Victor Hoffstein’s research about sleep apnoea concluded that sleep apnoea can worsen hypertension [1]. It is necessary to research and help clinicians about this matter. One way to do that is to research about sleep stages. This research aims to help people in the medical world to reduce the time required to classify the sleep stages from a patient who suffers from sleep disorders [2]. Before classify the sleep stages, clustering is needed. Clustering analysis plays an important role in scientific research and many applications. Clustering can group similar data into cluster, so in this case how many sleep stages of people with sleep apnoea can be known. This paper use K-Means Clustering method in MATLAB to kow the sleep stages that people with this sleep disorder may go through.

2. Methods

2.1. Obtaining raw data
In this study, it was used data from St. Vincent’s University Hospital. This database contains 25 full overnight polysomnograms with simultaneous three-channel Holter ECG, from adult subject with suspected sleep-disordered breathing. The signals recorded were EEG (C3-A2), EEG (C4-A1), left EOG, right EOG, submental EMG, ECG, oro-nasal airflow (thermistor), ribcage movements, abdomen
movements (Uncalibrated strain gauges), oxygen saturation (finger pulse oximeter), snoring (tracheal microphone), and body position.

2.2. **Preprocessing**

Preprocessing used to make all the input signals suitable for further processing method. Preprocessing consists of few steps as follows [3]

2.2.1. **Centering**. Shift the location of discrete amplitude distribution and makes its center locate the axis y=0, so the average amplitude of the signal becomes zero.

\[ \tilde{x} = x - \text{mean}(x) \]  \hspace{1cm} (1)

\( \tilde{x} \): new data  
\( x \): data

2.2.2. **Normalization**. Dividing each discrete amplitude value with the maximum amplitude value so the maximum amplitude of the signal balanced.

\[ v' = \frac{v - \text{min}_a}{\text{max}_a - \text{min}_a} \]  \hspace{1cm} (2)

2.3. **Feature Extraction**

Feature extraction is a process that aims at obtaining the characteristic from each voice signal obtained from the raw data (insert citation here). In this study, Feature Extraction is used to obtain characteristic from each data that has been going through the Preprocessing stage. In Feature Extraction itself, MFCC are used to process the power spectrum of the data. MFCC produces 24 parameters which consist of 12 Cepstral values and 12 first-order derivative values of these Cepstrals. Then, the output of this process will be devolved into frames, which each frame contains 28 feature values.

2.4. **K-Means Clustering**

K-means method are well known for its simple implementation and decent results for many practical applications in real life. The K-means method has been shown to be relatively effective in producing good clustering results. In this paper, K-means method are used to investigate or classify the sleep stages where the raw data that has been obtained before are preprocessed so that we can extract the feature from the raw data. Afterwards, the K-means clustering method uses to determine and grouping the data.

K-Means clustering is an algorithm to divide data into groups/clusters (K) build upon locations and the distance between input data. Each cluster has a centroid. Centroid is a middle value/mean value from a cluster. An object belongs to a cluster when it has the shortest distance to a centroid of the cluster itself [4].

In general, K-Means algorithm is as follows (1) Determining number of clusters (K); (2) Determining centroid, choose k starting points to become the initial values of the cluster centroids; (3) Grouping data based on the closest distance; (4) Recalculating the centroids, find the average of the cluster to become the new centroid, and (5) Repeat step 3 and 4 until the centroid is convergent.

Before the implementation of the clustering algorithm, the actual data samples are collected. Then give initial starting values for the centroids. Calculate distance of each data with the centroid of each cluster using squared Euclidean distance.
\[ d(x_j, c_j) = \sqrt{\sum_{i=1}^{n} (x_j - c_j)^2} \]  

\( d(x_j, c_j) \) : distance between j-th data and j-th centroid

j : index

c_j : j-th centroid

x_j : j-th data

n : numbers of data

Group the data into clusters based on the closest distance. Then calculate the mean of the data in each cluster to become the new centroid of each cluster. Repeat grouping and recalculating steps until each centroid is convergent [4,5].

3. Results and Discussion

3.1. K-Means Clustering and Silhouette Results

In this experiment, the purpose was to acquire the optimal number of clusters used in the experiment. An optimal number of clusters is when the experiment result has lesser number of negative values than the others. This experiment in this paper used 3 to 7 number of clusters, because 1 and 2 number of clusters cannot represent the overall sleep stages, as we know this experiment aims to find the sleep stages other than is already known which is wake and sleep, meanwhile the maximum number of clusters that processed is 7 due to resource constraints. In this k-means, it uses squared Euclidean distances and 100 iterations. K-means return an n-by-1 vector named idx containing the cluster indices of each point.

Silhouette used to see the cluster uniformity. If the silhouette image has an edge on the negative region, then the cluster not formed well because there is an outlier data included in the cluster. Otherwise, if all cluster does not have tails on negative territory, then the cluster results can be considered good enough. More positive the value is, the farther the distance between data in a cluster and the adjacent cluster, the value is 0 indicates that the distance between data in one cluster and its neighbours is very close and may even have same distance between two adjacent cluster, and the negative value means that the data in the cluster may on the wrong cluster.

K-means clustering and silhouette produce results in Table 1.

| Number of Clusters used | Number of Positive Values | Number of Negative Values |
|-------------------------|---------------------------|--------------------------|
| 3                       | 606762                    | 14235                    |
| 4                       | 595770                    | 25227                    |
| 5                       | 596624                    | 24373                    |
| 6                       | 593437                    | 27560                    |
| 7                       | 590875                    | 30122                    |
From Figure 1, it was used 3 clusters and the result showed that there were some negative values on the third cluster. Then, number of positive values were 607602 and number of the negative values were 14235 as noted on Table 1. From Figure 2, it was used 4 clusters and the result showed that there were some negative values between the third and the fourth cluster. Then, the number of positive values were 595770 and number of the negative values were 25227. The next experiment with another number of clusters still have some negative values as noted on Table 1 where on the number of clusters was 5, 6, and 7 each have 596624, 593437, 590875 of positive values and 24373, 27560, 30122 of negative values. With this result can be concluded that the result from the experiment with 3 number of clusters is the best result. Then, 3 clusters is the optimal number of clusters for this data, because it have less negative values which mean that it is more accurate than the others.

4. Conclusion
It is seen that every number of clusters which was experimented have negative values, but with 3 number of clusters, it have less negative values than the other number of clusters which mean that it was more accurate than the others. As described before, the negative values may mean that some data in that cluster was not formed well, it may went to the wrong cluster. Then, the experiment with 3 number of clusters is the optimal results, then this dataset can be divided into 3 sleep stages. Nevertheless, the results of
this experiment are not accurate because of the results of silhouette evaluation that has not perfect, there is still some negative value therefore this experiment still must be continued.

5. References
[1] Lavie P, Herer P and Hoffstein V 2000 Obstructive sleep apnoea syndrome as a risk factor for hypertension: population study BMJ 320 7233 479-482
[2] Aboalayon K A I, Faezipour M, Almuhammadi W S and Moslehpour S 2016 Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation Entropy 18 9 272
[3] Yulita I N, Houw Liong The and Adiwijaya 2012 Fuzzy Hidden Markov Models for Indonesian Speech Classification JACIII 16 3 381-387
[4] Wu J 2012 Advances in K-means clustering: a data mining thinking (Berlin: Springer Science & Business Media)
[5] Ghosh S and Dubey S K 2013 Comparative analysis of k-means and fuzzy c-means algorithms International Journal of Advanced Computer Science and Applications 4 4