ISSN: 1978-0524


# Clustering based feature selection using Partitioning Around Medoids (PAM)

Dewi Pramudi Ismi [a,1,*], Murinto [a,2]

[a] Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Indonesia
[1] *dewi.ismi@tif.uad.ac.id* [*]; [2] *murintokusno@tif.uad.ac.id*
** corresponding author*

ABSTRACT

High-dimensional data contains a large number of features. With many features, high dimensional data requires immense computational resources, including space and time. Several studies indicate that not all features of high dimensional data are relevant to classification result. Dimensionality reduction is inevitable and is required due to classifier performance improvement. Several dimensionality reduction techniques were carried out, including feature selection techniques and feature extraction techniques. Sequential forward feature selection and backward feature selection are feature selection using the greedy approach. The heuristics approach is also applied in feature selection, using the Genetic Algorithm, PSO, and Forest Optimization Algorithm. PCA is the most well-known feature extraction method. Besides, other methods such as multidimensional scaling and linear discriminant analysis. In this work, a different approach is applied to perform feature selection. Cluster analysis based feature selection using Partitioning Around Medoids (PAM) clustering is carried out. Our experiment results showed that classification accuracy gained when using feature vectors' medoids to represent the original dataset is high, above 80%.

*Keywords:*
Dimensionality reduction
Feature selection
Clustering
Partitioning Around Medoids (PAM)
High dimensional data

## I. Introduction

High-dimensional data contain a large number of features. A feature refers to a single measurable characteristic of the process being observed [1]. With many features, high dimensional data requires immense computational resources, including space and time. Several studies indicate that not all high dimensional data features are relevant to classification result. Thus removing irrelevant features of the original data increases classification accuracy [2]. High dimensional data may also include several correlated or redundant features to each other [3],[4], in which one can substitute the others. The curse of dimensionality is also introduced in the computation of high dimensional data. The data becomes sparse along with the increase of dimensionality. Sparse data cause accurate classification is hard to achieve. Therefore, dimensionality reduction is inevitable and required for the following reasons; improving classifiers' performance, reducing computation time and cost, and providing a better understanding of the data's underlying process [5].

There are two methods employed for dimensionality reduction purposes: feature selection and feature extraction [6]. Feature selection aims at finding a subset of features that are significant in predicting the classification output/class. Feature selection is selecting a minimum number of features that can be used to achieve high classification accuracy [6]. There are two different feature selection methods; exhaustive (deterministic) approach and heuristics (non-deterministic) approach. Each possible subset is used in wide feature selection to training the classifier, and the classification output is examined [7]. Backward feature selection and forward feature selection are examples of the exhaustive approach. In backward feature selection, the selection process is started by using the full features to train the classifier, and the classification result is recorded. The next steps in backward feature selection are removing one feature in each step, and the classification result of each step is evaluated. The removal of a feature in each step must increase classification accuracy [6]. The iteration continues until there is only one feature to be included in a subset. A feature is considered a relevant feature if the classification result is less accurate when it is not included in the feature subset. Forward feature selection works the other way around. An exhaustive approach needs to generate all possible subsets of features. An exhaustive approach is considered computationally prohibitive.

Many studies applied the heuristics approach for feature selection due to the high complexity of exhaustive feature selection methods. In the heuristics approach, feature selection becomes an optimization problem. The optimization algorithm is occupied with finding the best feature subset, which returns the highest classification accuracy. In [8], Particle Swarm Optimization (PSO) is used to generate the best feature combinations to construct the selected feature subset. Genetic Algorithm is also utilized in feature selection [9]. A recent study shows the implementation of another optimization algorithm, namely Forest Optimization Algorithm (FOA) for feature selection usage [10].

Another dimensionality reduction technique that has been proposed is feature extraction. Feature extraction aims to create a new feature set using linear and nonlinear combinations of the original features. Principal Component Analysis (PCA) is a famous example of a linear feature extraction method. In PCA, input data is mapped into a new space of smaller dimensions whereby the variance of input data in the new space is maximized [6]. Other feature extraction methods are factor analysis; they are multidimensional scaling [11], linear discriminant analysis [12], and locally linear embedding [13].

In this work, we try to use a different approach to perform feature selection. We use the clustering method to generate a subset of features to be fed into a classifier. Clustering is performed to group feature vectors such that similar feature vectors are located in the same cluster. It is assumed that the cluster centers resulting from the clustering process can replace the original features. The Centre of a cluster represents the other feature vectors within that cluster. Partitioning Around Medoids (PAM) clustering algorithm is chosen for this feature selection purpose. Our proposed method is wrapper feature selection; hence the evaluation is based on the final classification accuracy.

## II. Literature Review

### *A. Feature Selection*

The performance of a classifier is greatly affected by the size of the input data. With its enormous number of features/dimensions, high dimensional data increase the computation complexity of classifiers. Thus, increasing the classifier's performance that working on high dimensional data becomes the objective of many research works. Dimensionality reduction is an attempt to increase the performance of classifiers by diminishing the size of data. Ideally, classifiers should have the ability to distinguish important features and irrelevant ones [6]. However, there are several reasons why dimensionality reduction is conducted as a separated process:

a. Decreasing data dimensionality contributes to increasing the performance of classifiers during the training phase and contributes to increasing the performance of classifiers during the testing phase.
b. When a feature is considered to be 'unnecessary' for class prediction, the cost of processing this feature during training is such a waste.
c. High dimensional data give much chance to overfitting problem. Small data usually leads to a simpler model, and a simpler model tends to generalize better.
d. Small data explains a better idea about the process that underlies the data.

Two different ways of diminishing data dimensionality are feature selection and feature extraction. Feature selection focuses on finding a subset of the original features relevant to classification results. The greedy approach is implemented for feature selection, namely sequential forward feature selection, and sequential backward feature selection. The heuristics approach also has been implemented for feature selection, such as the Genetic Algorithm, Particle Swarm Optimization, and Forest Optimization Algorithm. Feature extraction maps the original dataset into some other spaces which have less dimension. The most well-known feature extraction method is Principal Component Analysis (PCA). Other than PCA, several feature extraction methods are linear discriminant analysis, multidimensional scaling, isomap, etc.

### *B. Partitioning Around Medoids (PAM)*

Partitioning Around Medoids (PAM) is a clustering algorithm in which the k-medoids paradigm is applied. It was proposed in 1987 by Kaufman and Rousseeuw. Partitioning around Medoids is considered partitional clustering, similar to k-means clustering. Unlike k-means clustering, which uses the mean of the data points within a cluster to become cluster center, PAM clustering uses data point, which has a less total distance of the resultant clustering.

It starts from an initial set of medoids and iteratively replaces one of the medoids with one of the non-medoids if it improves the resultant clustering's total distance. It selects k representative medoid data items arbitrarily. The total swapping cost S is calculated for each pair of non-medoid data item x and selected medoid m. If S< 0, m is replaced by x. After that, each remaining data item is assigned to a cluster based on the most similar representative medoid. This process is repeated until there is no change in medoids.

Partitioning Around Medoids Algorithm

1. Identify the number of clusters *k*
2. Select *k* random data points as medoids
3. For each pair of non-medoid data point xi and selected medoid m**k**, calculate the total swapping cost S(xi, m**k**). For each pair of xi and m**k,** if S < 0, m**k** is replaced by xi
4. Assign each data point to the cluster with the nearest medoid
5. Repeat steps 2-3 until there is no change in the medoids.

For each non-medoid data point, swapping cost is calculated by subtracting its distance to the new centroid candidate from its distance to the current centroid of the cluster it belongs to. Its cumulative swapping cost evaluates each new centroid candidate. If a new centroid candidate's swapping cost is less than 0, this new centroid candidate is selected to replace the current centroid.

PAM algorithm complexity to calculate cost function S in each iteration (step 3) is $O(k(n-k)^2)$. Moreover, the PAM algorithm complexity to recalculate the entire cost function is $O(n^2k^2)$.

## III. Method

A classification process is usually done by feeding data into a classifier, and classes of the data have resulted. This way is illustrated by Fig. 1.

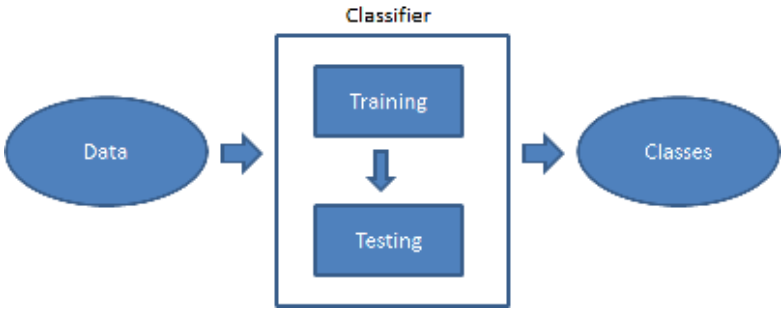


Fig. 1. Classification steps without dimensionality reduction

In this research work, an additional step is added prior to feeding the data into the classifier, that is, to perform feature selection of high dimensional data. This additional step intends to reduce the dimensionality of the data by selecting a feature subset that can represent the whole feature set. The steps performed in this work are illustrated in Fig. 2.

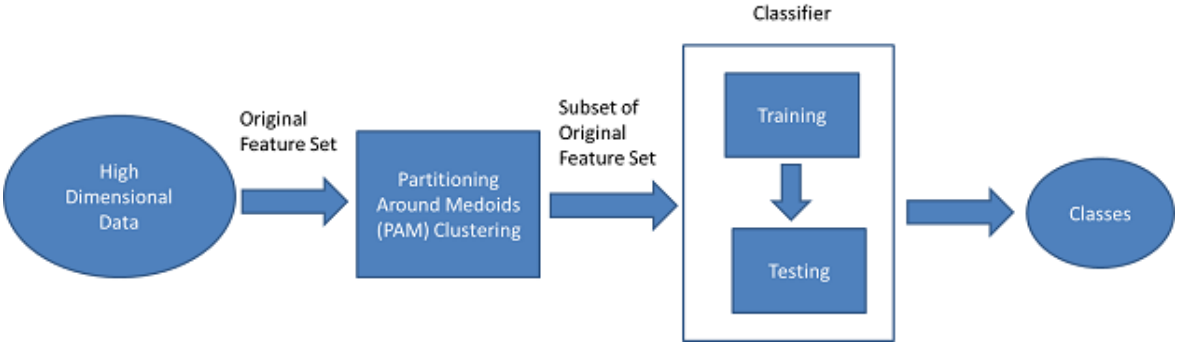


Fig. 2. Steps of PAM based feature selection method

### *A. High Dimensional Dataset Collection*

In this work, two public datasets taken from [14] are used; Human Activity Recognition (HAR) Dataset [15][16][17] and Multiple Features Dataset [18][19].

1. Human Activity Recognition (HAR) Dataset
   This dataset consists of 30 volunteers performing six daily activities such as walking, walking upstairs, walking downstairs, standing, sitting, and laying. A smartphone is attached in their waists. This dataset consists of 561 features, including triaxial acceleration from the accelerometer (total acceleration), the estimated body acceleration, and triaxial angular velocity from the gyroscope. These properties are measured with time and frequency domain variables.
2. Multiple Features Dataset
   This dataset consists of features of handwritten numerals (`0'--`9'). There are 200 patterns for each numeral (a total of 2,000 patterns) that have been digitized in binary images. These digits are represented the following six feature sets:
   a. mfeat-fou: 76 Fourier coefficients of the character shapes;
   b. mfeat-fac: 216 profile correlations;
   c. mfeat-kar: 64 Karhunen-Love coefficients;
   d. mfeat-pix: 240 pixel averages in 2 x 3 windows;
   e. mfeat-zer: 47 Zernike moments;
   f. mfeat-mor: 6 morphological features.

### *B. Partitioning Around Medoids (PAM) Clustering*

In this work, two public datasets taken from [14] are used; Human Activity Recognition (HAR) Dataset [15][16][17] and Multiple Features Dataset [18][19].

Partitioning Around Medoids (PAM) clustering is performed onto the original HAR dataset and the original multiple features dataset. Assuming that original dataset *X* consisting of *N* data and *D* features (attributes) is used. The dataset can be formulated as follows.

$$X = \{(X_{11}, X_{12}, X_{13}, \dots, X_{1D}),$$
$$(X_{21}, X_{22}, X_{23}, \dots, X_{2D}),$$
$$\vdots$$
$$(X_{N1}, X_{N2}, X_{N3}, \dots, X_{ND})\}$$

The feature vector of dataset *X*, *Fi*, is defined as the vector consisting of the $i^{th}$ attribute values. Hence, *D* numbers of different feature vectors inferred from dataset *X*, namely *F1, F2, F3, …, FD*. *Fi* contains *Xji* where j is started from 1 and is ended at *N*, while *i* denotes the attribute order.

$$F_1 = (X_{11}, X_{21}, X_{31}, \dots, X_{N1}),$$
$$F_2 = (X_{12}, X_{22}, X_{32}, \dots, X_{N2}),$$
$$\vdots$$
$$F_D = (X_{1D}, X_{2D}, X_{3D}, \dots, X_{ND})\}$$

PAM clustering algorithm is applied in this work, aiming at reducing the dimension of dataset *X*, namely *D*. It is done by selecting *k* different feature vectors, whereby $k < D$, which can represent the whole feature vectors. The selection of *k* feature vectors out of *D* feature vectors is performed through the following steps.

1. Feature vectors of dataset X, namely F1, F2, F3, ... FD are used as input of the PAM clustering algorithm.
2. PAM clustering algorithm is performed to divide the input (feature vectors) into groups based on similarity.
3. PAM clustering algorithm produces *k* clusters of feature vectors, and each cluster has a cluster center (centroid).
4. The cluster center (centroid) is the medoid of the cluster and is a feature vector. The cluster center is then used to represent members of the cluster.
5. Finally, *k* feature vectors (which are centroids of produced clusters) represent *D* feature vectors of the original dataset *X*.

These *k* feature vectors are used to replace the original dataset when training the classifier and test the classifier's performance. Moreover, the number of clusters *k* becomes the new dimension of the reduced dataset. An example of this dimensionality reduction technique is illustrated in Fig. 3.

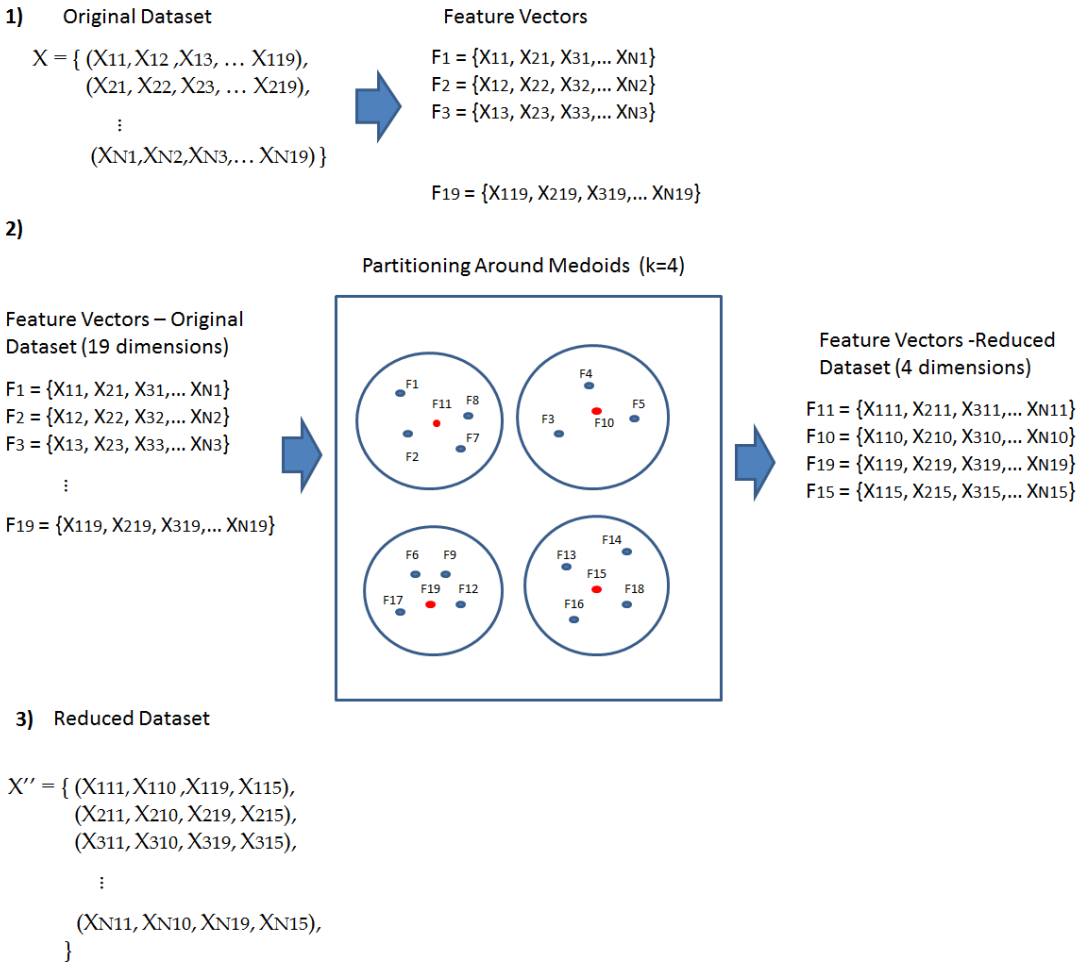


Fig. 3. The original dataset consisting of 19 features ($F_1$ to $F_{19}$) is reduced into four features ($F_{11}$, $F_{10}$, $F_{19}$, $F_{15}$)

### *C. Classification using Naïve Bayes Classifier*

The reduced dataset produced from the previous step is then fed into the Naïve Bayes classifier. The dataset is split into training data and testing data. Then, the accuracy of classification is recorded. This reduced dataset is used to train the Naïve Bayes classifier and test the classifier through 10 folds cross-validation method. The accuracy gained by this reduced dataset is analyzed and compared with the original dataset classification accuracy. Naive Bayes Classifier is used as it is highly scalable with the number of features and is not sensitive to irrelevant features.

## IV. Results and Discussion

### *A. Method*

This work used the PAM algorithm to select a subset of features. This subset of features consists of medoids data. Medoids data generated by PAM are used as representations of the full features of the dataset. The number of clusters ($k$) generated in the PAM algorithm has to be defined. We experimentally employed various number of clusters; 50 ,60, 70, 80, 90, 100, 150, 200, 250, 300. Compared to using PCA, which automatically finds the number of feature subset to be generated, PCA computational complexity is dependent on the number of data and number of features of the dataset ($O(D^2N + D^3)$ where D is a dimension, N is the number of data). If the dataset has a high dimension, it causes a heavy computation workload. On the other side, PAM computational complexity is $O(N^2k^2)$. Since k is set to be less than $D$ of the dataset, using PAM to perform feature selection can reduce the computation cost if it has a large dimension. This experiment showed that the number of features returning best classification accuracy in both datasets is 100 features (HAR Dataset) and 200 features (Multiple Features Dataset). It means less than 50% of the total number of features in both datasets.

### *B. Class Prediction Correctness of HAR Dataset*

The experiment result showed that feature selection using PAM (Partitioning Around Medoids) on the HAR dataset effectively increased the classification accuracy. Table 1 showed that all subset of features used to train and test the classifier returned higher than 80% of classification accuracy. The best classification accuracy is achieved when 100 features are used. In contrast, the original dataset returned only 76.77 % of classification accuracy (shown in Table 3). This result explains that reducing the dimensionality of data leads to enhance classifier performance. This result also explains that the original HAR dataset contains many irrelevant features that do not contribute to classification accuracy. Fig. 4 shows that all attempted feature subsets' classification accuracy is higher than the original dataset's classification result.

Table 1. Performance of Naive Bayes Classifier on Reduced HAR Dataset

| | Number of Features | Correct Class Prediction (%) | Error Class Prediction (%) |
|---|---|---|---|
| HAR Dataset | 50 | 86.80 | 13.20 |
| | 60 | 87.01 | 12.99 |
| | 70 | 87.13 | 12.87 |
| | 80 | 87.41 | 12.59 |
| | 90 | 86.58 | 13.42 |
| | **100** | **87.78** | **12.22** |
| | 150 | 87.51 | 12.49 |
| | 200 | 87.71 | 12.29 |
| | 250 | 87.11 | 12.89 |
| | 300 | 85.36 | 14.64 |

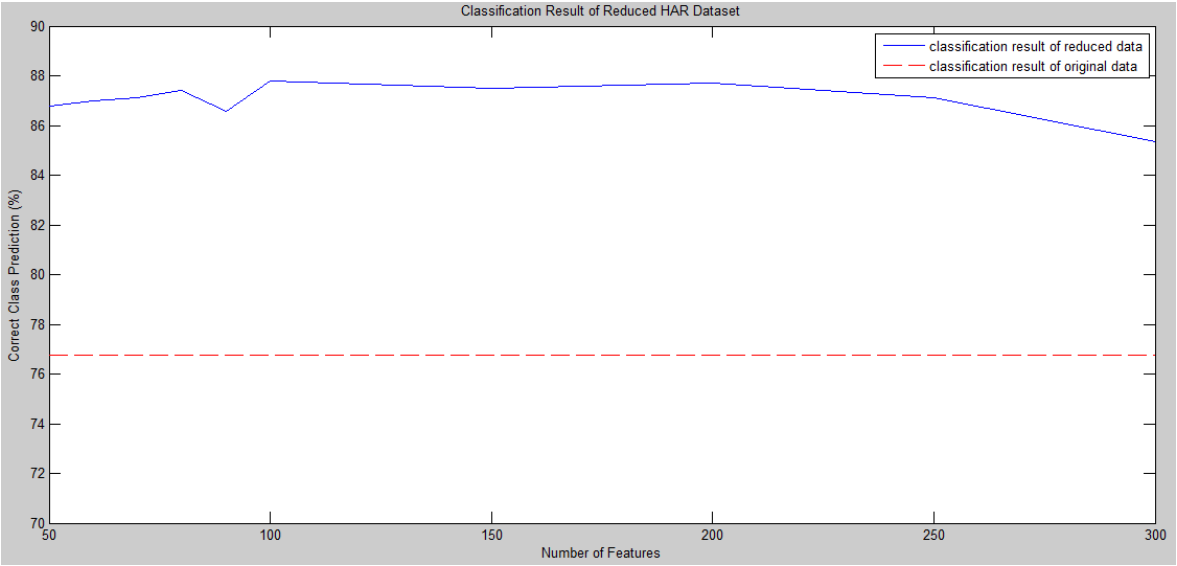


Fig. 4. Classification Result of Reduced HAR Dataset

### *C. Class Prediction Correctness of Multiple Features Dataset*

The experiment result showed that feature selection using PAM (Partitioning Around Medoids) on Multiple Features dataset returned higher than 90% classification accuracy (Table 2). The best classification accuracy is achieved when 200 features are used. If compared to the original dataset, classification using a reduced dataset resulted in lower classification accuracy. As shown in Table 3, the original Multiple Feature dataset produced 95.35% classification accuracy. This result explains that each feature of the Multiple Feature dataset contributes significantly to class prediction, such that reducing the dimensionality of the data does not improve classifier accuracy. However, the high classification accuracy of the reduced dataset (> 90%) is still achieved. Fig. 5 shows a comparison of classification accuracy between the reduced dataset and the original dataset.

Table 2. Accuracy Values in Testing Data After Tunning Hyper Parameter

| | Number of Features | Correct Class Prediction (%) | Error Class Prediction (%) |
|---|---|---|---|
| Multiple Features Dataset | 50 | 91.80 | 8.20 |
| | 60 | 92.45 | 7.55 |
| | 70 | 92.25 | 7.75 |
| | 80 | 92.30 | 7.70 |
| | 90 | 92.10 | 7.90 |
| | 100 | 92.70 | 7.30 |
| | 150 | 91.90 | 8.10 |
| | **200** | **93.95** | 6.05 |
| | 250 | 93.90 | 6.10 |
| | 300 | 93.85 | 6.15 |

Table 3. Classification Accuracy on Original Dataset

| Dataset | Correct Class Prediction (%) | Error Class Prediction (%) |
|---|---|---|
| HAR | 76.77 | 23.23 |
| Multiple Features | 95.35 | 4.65 |

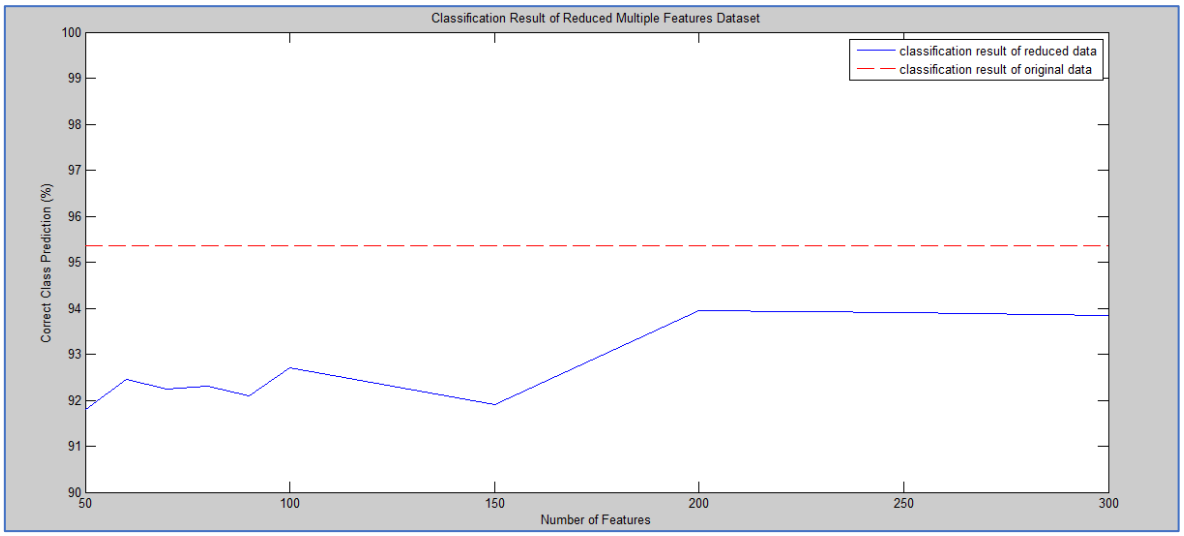


Fig. 5. Classification result of reduced multiple feature dataset

## V. Conclusion

Feature selection using a clustering approach, especially Partitioning Around Medoids (PAM), had been performed in this work. The experiment results explain several conclusions. A set of medoids produced by Partitioning Around Medoids (PAM) clustering applied on the original dataset can represent the original dataset. Thus, occupying this set of medoids onto classifiers produced high classification results. Our experiments showed higher than 80% classification accuracy when using a reduced dataset consisting of medoids. This result explained that reducing the dimensionality of the dataset using the clustering approach, namely to use cluster centers to represent feature vectors, effectively diminished irrelevant features that do not contribute significantly to classification results. Further research work can be performed to enhance clustering-based feature selection methods, such as applying clustering-based feature selection on 3-dimensional data and applying different clustering methods such as hierarchical clustering, CLARANS, and DBSCAN feature selection purpose.

## References

[1] G. Chandrashekar, F. Sahin. A survey on feature selection methods. Computers and Electrical Engineering. Vol. 40. Issue 1. January 2014. pp 16-28.

[2] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu. Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection. IEEE Transactions on Knowledge and Data Engineering. Vol. 26. Issue 9. September 2014. pp. 2138-2150.

[3] R. Duda, P. Hart, and D. Stork. Pattern Recognition. 2nd ed. New York, NY, USA: Wiley, 2001.

[4] H. Liu, X. Wu, and S. Zhang, "Feature selection using hierarchical feature clustering," in Proc. ACM Int. Conf. Inform. Knowl. Manage., New York, NY, USA, 2011.

[5] V.B. Canedo., N.S. Marono, A.A. Betanzos. A review of feature selection methods on synthetic data. Knowledge and Information System (2013) 34:483.

[6] E. Alpaydin, Introduction to Machine Learning 2nd edition, MIT Press, 2006.

[7] T.M. Cover, J.V.P. Campenhout. On the Possible Orderings in the Measurement Selection Problem. IEEE Transactions on Systems, Man, and Cybernatics. Vol. 7. Issue 9. September 1977. pp. 657-661.

[8] X.Wang, J.Yang, X.Teng, W.Xia, R. Jensen. Feature Selection based on Rough Sets and Particle Swarm Optimization. Pattern Recognition Letters. Vol 28. Issue 4. March 2007. pp. 459-471.

[9] C.L. Huang, C.J. Wang. A GA-based feature selection and parameters optimization. Expert Systems with Applications. Vol 31. Issue 2. August 2006. pp 231-240.

[10] M. Ghaemi, M.R.F Derakhshi. Feature selection using Forest Optimization Algorithm. Pattern Recognition.Vol. 60. December 2016. pp. 121-129.

[11] T.F. Cox, M.A.A. Cox. Multidimensional Scaling. London: Chapman and Hall.1994

[12] G.J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. New York: Willey. 1992

[13] S.T. Roweis, L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. Vol.290. December 2000. pp: 2323-2326.

[14] UCI Machine Learning Repositories http:// http://archive.ics.uci.edu/ml/

[15] D. Anguita, A. Ghio, L. Oneto, X. Parra and J.L.R.Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012.

[16] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L.R.Ortiz. Energy Efficient Smartphone-Based Activity Recognition using Fixed-Point Arithmetic. Journal of Universal Computer Science. Special Issue in Ambient Assisted Living: Home Care. Volume 19. Issue 9. May 2013.

[17] J.L.R. Ortiz, A. Ghio, X. Parra, D. Anguita, J. Cabestany, A. Catala. Human Activity and Motion Disorder Recognition: Towards Smarter Interactive Cognitive Environments. 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium, 24-26 April 2013.

[18] M. van Breukelen, R.P.W. Duin, D.M.J. Tax, J.E. den Hartog, Handwritten digit recognition by combined classifiers, Kybernetika. vol. 34. no. 4. 1998. pp 381-386.

[19] M. van Breukelen, R.P.W. Duin. Neural Network Initialization by Combined Classifiers, in: A.K. Jain, S. Venkatesh, B.C. Lovell (eds.). ICPR'98. Proc. 14th Int. Conference on Pattern Recognition (Brisbane, Aug. 16-20).1998.