# The Silhouette Width Criterion for Clustering and Association Mining to Select Image Features

Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Image data are normally unstructured and high dimensional due to the photography technology advancement such that an image can be taken at a wide range of resolution levels. To overcome such problem, data miners may consider selecting only a minimal set of features that are really important for classifying their images. Feature selection is a popular method for reducing dimensions in data. However, most feature selection algorithms return results in form of score for each feature. It is still difficult for data miners to choose features based on such scoring scheme because they may not know which score range is the best for their data classification at hand. Therefore, in this research, we aim to assist data miners and novice data analysts on solving dimensionality problem by finding for them the best optimal set of features, instead of just reporting the scores of all features and leaving the selection step to be the burden of miners. We select optimal set of features by firstly apply clustering technique to group similar features based on their scores. We thus propose the silhouette width criterion for selecting the optimal number of clusters during the cluster analysis step. After that we perform association mining to analyze relationships that may exist among different subsets of features toward the target attribute. Our method finally reports user the best subset of features to be potentially used further for data classification. We demonstrate performance of our proposed method on the satellite forest image data in Japan.**

***Index Terms*—Image data, feature selection, clustering, silhouette criterion, forrest type classification.**

## I. Introduction

With the rapid development of current electronic devices such as sensors and cameras, the outputs from these devices are of high quality and also high dimensions. Unfortunately, high dimensionality is still an unsolvable issue for many existing data mining and machine learning algorithms. Data with overwhelming attributes or dimensions can be a major cause of low computational performance. It can be even worse when such data may cause a creation of classifying model with low predictive accuracy due to the search for discriminative set of features is obscured by so many irrelevant features. Most classification algorithms are not designed to efficiently handle such high dimensionality problem.

Therefore, the numerous feature selection techniques have been proposed as a pre-classification step for solving the high dimensionality problem. Several research teams introduce many ways to reduce the number of features. The reduced set of features has been proven experimentally increasing the performance of learning process and also being able to build an accurate classification model. Generally, feature selection techniques can be divided into three classes [1]. The first class is called filter method, such as CfsSubsetEval [2], Information Gain, and Chi-Square [3]. The second class is the wrapper method [4], [5]. The filter method introduces some form of scoring computation without actually building a model, whereas the wrapper approach scoring the selected set of features by observing the error made by the classifying model. The last class is called the embedded method; it combines the advantages from both the filter and wrapper methods [5].

Xie *et al.* [6] proposed the association rule mining technique to calculate weight for find the optimal features that are closely correlative with the class attribute, but the proposed technique is quite complex and performance test with cross validation. Nuntawut *et al*. [7] proposed the filter method for feature selection based on association rule mining such that the specific set of association rules that the rules' consequence is the target class. But this feature selection algorithm does not work automatically because human is the one who select the features one by one based on the feature scores reported from the algorithm. Therefore, Nuntawut *et al*. [8] improved the algorithm by proposing clustering technique to cluster the feature scores to assist users on finding an appropriate groups of features. The clustering process is supposed to be automatic in the sense that the number of clusters should be judged by the process itself. However, the clustering algorithm is still semi-automatic in the sense that users must specify the suitable number of feature clusters.

This research, thus, aims at extending the previous work of Nuntawut *et al*. [7], [8] by proposing a silhouette width criterion for automatic setting of initial cluster numbers. We also add confidence criteria into feature selection based on association rule mining technique to increase performance. Experimental results confirm the efficacy of our proposed method that can extract only relevant set of image features from ASTER satellite resulting in better recognition for each forest type.

## II. Materials and Methods

### A. Feature Selection Based on Association Rule Mining

Association rule mining is finding the frequent patterns in

Manuscript received September 20, 2017; revised January 10, 2017. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology (SUT), Nakhon Ratchasima 30000, Thailand (corresponding author: N. Kaoungku; Tel: +66872155059; e-mail: nuntawut@sut.ac.th, mikaiterng@gmail.com, arc_angle@hotmail.com, kittisakThailand@gmail.com, nittaya@sut.ac.th).

doi: 10.18178/ijmlc.2018.8.1.665

database and present them in the form of association rules [9]. Generally, there can be so many possible association rules from this technique. Therefore, some constraints are necessary for reducing such exponential growth. There are two popular criteria: support and confidence. Support is the frequency of the occurring event, as shown in (1). Confidence is the proportion of frequency of co-occurring events to the frequency of antecedent event, as shown in (2).

$$\text{Support}, \text{Supp}(X \to Y) = P(X \wedge Y) \tag{1}$$

$$\text{Confidence}, \text{Conf}(X \to Y) = \frac{\text{Supp}(X \to Y)}{\text{Supp}(X)} \tag{2}$$

This technique had been successfully applied to multiple disciplines such as marketing to increase sales. Nuntawut *et al*. [7] applied this technique to find optimal feature set from high dimensional dataset by finding association rules that the target class appears in the consequence of the rule. Then, consider the features or attributes that are most influencing the target class. The algorithm consists of 4 steps:

Step 1: define minimum frequency threshold, support, and confidence. Find frequent patterns and then generate association rules based on the Apriori algorithm [10].

Step 2: select only association rules that their consequence is target class.

Step 3: count features that appear on association rules.

Step 4: calculate frequent features in percentage, as in (3). Then, remove any feature having percentage of frequency appearance in the set of association rules lower than the specified minimum frequency threshold.

$$\text{FrequentFeature} = \frac{\text{AppearFrequency}}{\#\text{Rules}} \times 100 \tag{3}$$

### B. k-Means Clustering

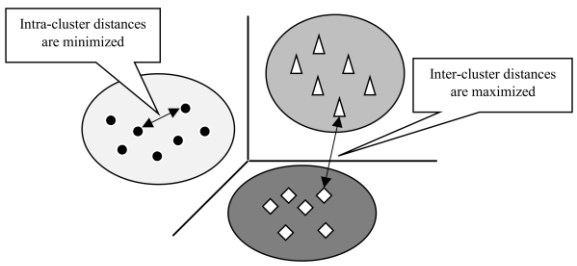


Fig. 1. Forming three clusters with minimized intra-cluster distance but maximized the inter-cluster distance.

k-Means algorithm is unsupervised learning method to form data into clusters based on data similarity regardless of the target class information. Fig. 1 depicts the idea of clustering such that distance between data in the same cluster (intra-cluster) is low, whereas the distance between data in one cluster to data in another cluster (inter-cluster) is high [11], [12]. The k-means algorithm can be explained by the following 4 steps:

Step 1: define the number of clusters (K) and randomly pick k instances as the initial cluster centroids.

Step 2: assign all data points to the closest centroid by measuring the distance such as the Euclidean distance.

Step 3: re-compute the centroid of each cluster by calculating mean value of all the data points in the cluster.

Step 4: repeat steps 2 and 3 until the centroid does not change.

### C. Silhouette Coefficient

The shortcoming of k-means clustering is the appropriate choice of *k*, which is the number of clusters. Silhouette coefficient is a popular measure for considering such parameter. The silhouette coefficient can be computed by using average distance between data points in the same cluster compared against average distances between data points in other clusters. Fig. 2 shows main concept of the silhouette coefficient to calculate the silhouette average of all cluster.

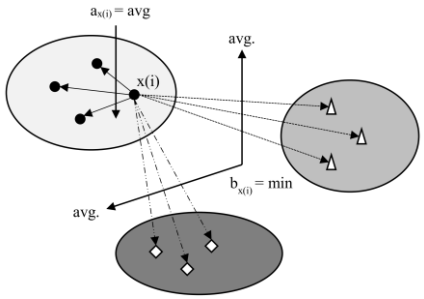


Fig. 2. Concept of the silhouette coefficient.

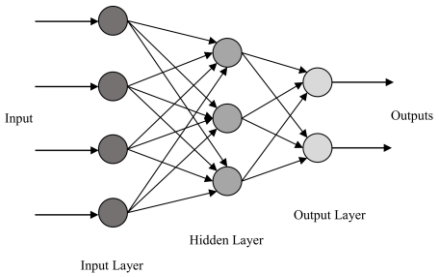


Fig. 3. The architecture of artificial neural networks.

Define $K$ to be cluster composing of data $x(i)$ and $a_{x(i)}$ is average distance between $x_{(i)}$ to every data point in the cluster $K$. The notation $b_x$ is minimum average distance between $x(i)$ and every data point in other clusters that are not a member in $K$. The calculation [13] of the silhouette coefficient of $x(i)$, the silhouette average of each cluster, and the silhouette average of all cluster can be shown as in (4), (5), and (6), respectively.

$$S_{x(i)} = \frac{b_{x(i)} - a_{x(i)}}{\max(a_{x(i)}, b_{x(i)})} \tag{4}$$

where

$x(i)$ = data point in the cluster, $i$ = 1, 2, 3, …, $n$,

$a_{x(i)}$ = average distance between $x_i$ and every data point in the same cluster, and

$b_{x(i)}$ = minimum average distance between $x_i$ and every data point in other clusters.

$$S_k = \frac{1}{n}\sum_{i=1}^{n} S_{x(i)} \tag{5}$$

where $k$ = number of clusters, and

$n$ = number of data points in the same cluster.

$$S_{avg} = \frac{1}{m}\sum_{k=1}^{m} S_k \tag{6}$$

where $m$ is number of all clusters.

### D. Artificial Neural Networks

Artificial neural networks is a simulation of human brain with computer program that can self-adjusting from learning the input values. The remarkable feature of this technique is that it consists of many nodes in the hidden layer in which parallel connections are effective for data classification [14]. Fig. 3 shows general architecture of artificial neural networks consisting of nodes and edges between nodes. Form the figure, the network can be partitioned based on node layout into 3 layers. The first layer is input layer; the second is hidden layer (this layer can have more than 1 layer), and the final layer is output layer.

## III. Proposed Work

In this section, we present the proposed process of silhouette width criterion consideration for automatic clustering of feature sets with the main focus of finding optimal feature to be discovered by association rule mining. The idea is that we use the silhouette coefficient to find the appropriate number of clusters for clustering the feature scores from feature selection obtained from the association rule mining. The objectives are to increase the predictive accuracy and to reduce the data dimensions of forest type dataset.

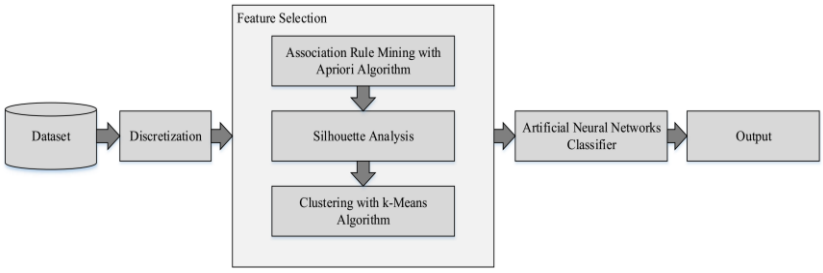


Fig. 4. The concept of feature selection based on association rule mining.

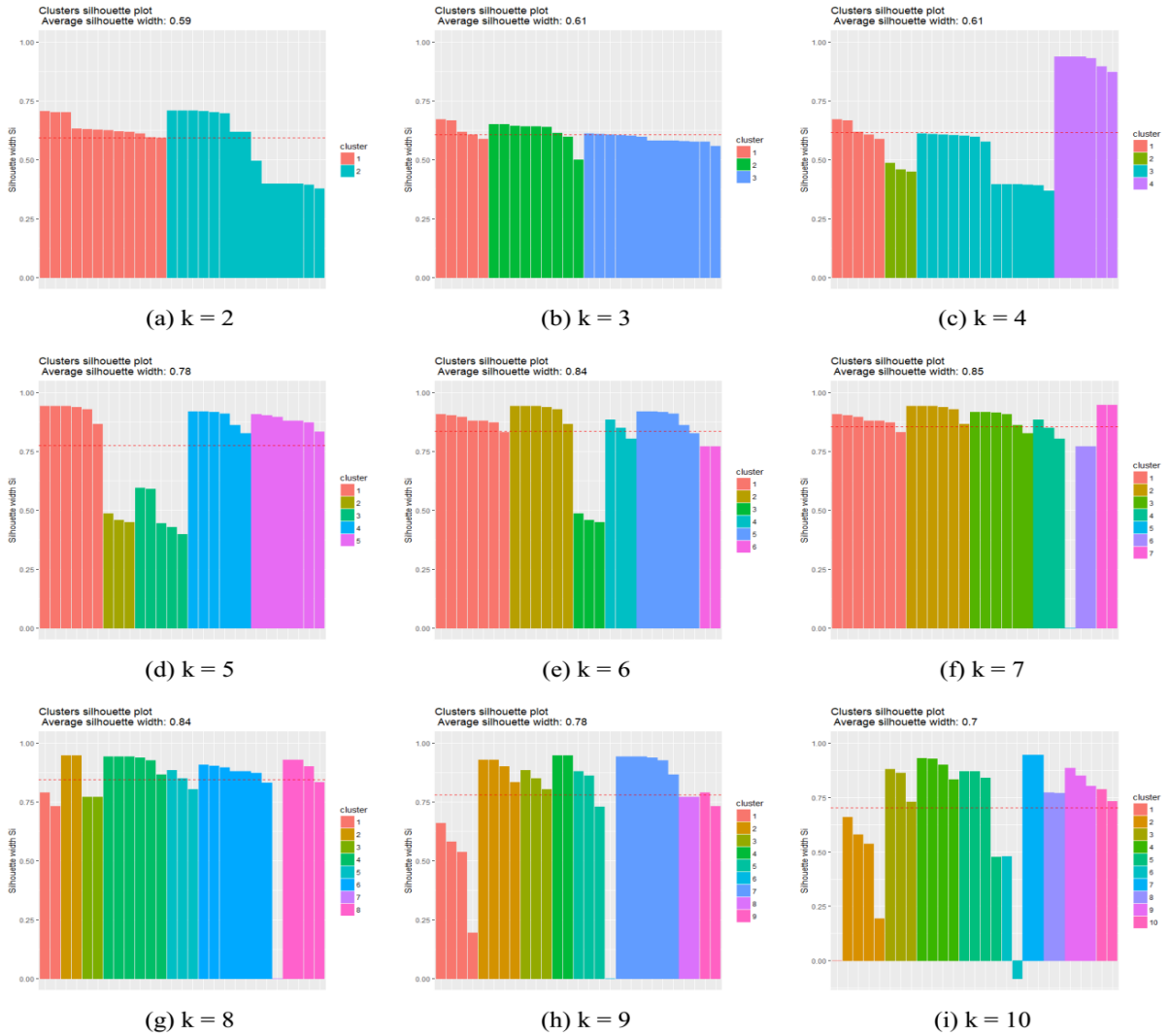


Fig. 5. Comparative graphs showing average silhouette widths of different cluster numbers.

Fig 4 shows the steps of the proposed process, which consists of three phases: phase 1, read the dataset from file or database and then perform discretization with chi-square algorithm if the data type is numeric. This discretization step is necessary for association rule mining that can handle only categorical values. Phase 2 is the feature selection method that

consists of three steps:

Step 1: find frequent patterns and generate from these patterns association rules in a format "*IF condition THEN consequence.*" This step is done through Apriori algorithm with initial 2 thresholds: support and confidence. We also constrain the algorithm to generate rules with target class appeared in the consequence part of the rule. The result from this step is a set of features with scores computed as feature frequency in association rules and average confidence of each feature.

Step 2: cluster the features based on their scores with different number of clusters. For each number of clusters, calculate the silhouette coefficient to find the best number of clusters. The higher silhouette coefficient means the better formation of clusters. The result in this step is the optimal parameter *k* to be used in the k-means clustering on step 3.

Step 3: perform k-means clustering with the initial number of cluster (k) according to the recommend value from step 2. We then select a set of features from a cluster showing mean confidence higher than other clusters. The result in this step is optimal feature set to be used for classification.

Finally, phase 3 is the building of classifier using artificial neural networks.

TABLE I: COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY, NUMBER OF FEATURES, AND AVERAGE SILHOUETTE WIDTH

| Number of Clusters (k) | Accuracy | Number of Features | Average Silhouette Width |
|---|---|---|---|
| 2 | 80.31% | 20 | 0.59 |
| 3 | 81.54% | 14 | 0.61 |
| 4 | 82.77% | 11 | 0.61 |
| 5 | 82.77% | 11 | 0.78 |
| 6 | 82.77% | 11 | 0.84 |
| **7** | **84.62%** | 10 | **0.85** |
| 8 | 80.00% | 7 | 0.84 |
| 9 | 79.69% | 5 | 0.78 |
| 10 | 78.46% | **3** | 0.70 |

## IV. EXPERIMENTAL RESULTS

To test performance of the proposed method of feature selection based on the silhouette width criterion for clustering relevant featured discovered by association rule mining, we use the forest type with high-resolution imaging from ASTER satellite that has been publicly available at the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). The data are divided into training dataset (198 instances) and test dataset (325 instances). We initialize Apriori algorithm to discover feature sets with support = 0.1 and confidence = 0.1. We experiment with number of clusters (k) between 2 to 10 clusters.

Table I shows comparative results of classification accuracy, number of features, and average silhouette. Fig. 5 shows comparative average silhouette widths of different clusters. It can be seen that when the number of cluster = 7, the average silhouette coefficient is maximized (0.85). At this maximum coefficient value, the predictive accuracy is as high as 84.62%. Moreover, the number of features can be reduced from 26 down to 10. Characteristic of number of features according to the changing number of clusters has been captured and shown in Fig. 6.

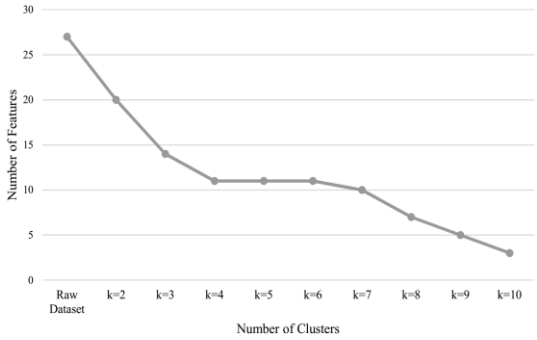


Fig. 6. The effect of cluster numbers to the number of features.

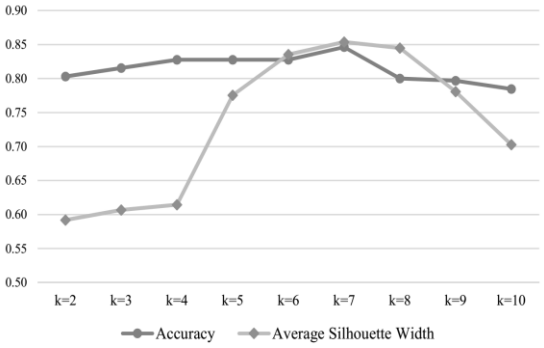


Fig. 7. The accuracy and average silhouette width characteristics.

Fig. 7 shows the comparisons of predictive accuracy and average silhouette width as the number of clusters has been varied from 2 to 10. It can be seen that the average silhouette has direct and positive impact to the classification accuracy. This is observed from the graph that when the silhouette width is low, the accuracy is also low. When the silhouette width is high, the accuracy is high as well.

## V. CONCLUSION

This research aims at studying a novel method to use silhouette width criterion for cluster analysis with the main focus of finding optimal feature set to be used for building classification model. Set of features are discovered with the association rule mining method. The proposed feature subset selection method is to be applied on classifying data with high dimensionality such as satellite image data. Our proposed method works with three main phases. Firstly, find and score relevant set of features based on association rule mining technique. Secondly, apply silhouette width criterion to find optimal parameter *k* for the next phase of feature clustering and add average confidence threshold of each cluster to feature score for increasing clustering performance. From the experimental results, we can conclude that the proposed method can select a discriminative set of features resulting in a highly accurate classification model.

## REFERENCES

[1] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 102-118, 2008.

[2] Z. N. Hamilton, "Correlation-based feature subset selection for machine learning," Ph.D. Dissertation, Department of Computer Science, Waikato University, 1998.

[3] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," in *Proc. International Workshop on Data Mining for Biomedical Applications*, 2006, pp. 106-115.
[4] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, pp. 1205-1224, 2004.
[5] Y. Saeys, P. Rouzé, and Y. Van de Peer, "In search of the small ones: Improved prediction of short exons in vertebrates, plants, fungi and protists," *Bioinformatics*, vol. 23, no. 4, pp. 414-420, 2007.
[6] J. Xie, J. Wu, and Q. Qian, "Feature selection algorithm based on association rules mining method," *In Proc. the ACIS International Conference on Computer and Information Science*, pp. 357-3622009.
[7] N. Kaoungku, K. Suksut, R. Chanklan, K. Kerdprasop, and N. Kerdprasop, "Data classification based on feature selection with association rule mining," *The 25th Int. MultiConference of Engineers and Computer Scientists (IMECS2017)*, Hong Kong, China, 15-17 March 2017, pp. 321-326.
[8] N. Kaoungku, K. Kerdprasop, and N. Kerdprasop, "A method to clustering the feature ranking on data classification using an ensemble feature selection," *International Journal of Future Computer and Communication*, vol. 6, no. 3, September, pp. 81-85, 2017.
[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216, 1993.
[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. The 20th Int. Conf. on Very Large Data Bases (VLDB)*, 1994, vol. 1215, pp. 487-499.
[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.
[12] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.
[13] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," in *Proc. IEEE Int. Conf. on Computational Intelligence and Multimedia Applications*, 2007, vol. 2, pp. 13-17.
[14] B. Yegnanarayana, *Artificial Neural Networks*, PHI Learning Pvt. Ltd., 2009.

**N. Kaoungku** is currently a lecturer at School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. He received his doctoral degree, master degree, and bachelor degree in computer engineering from SUT, in 2015, 2013, and 2012, respectively. His current research includes data mining, knowledge engineering, and semantic web.

**K. Suksut** is currently a doctoral student with the School of Computer Engineering, SUT, Thailand. He received his bachelor degree in computer engineering from SUT in 2011, master degree in computer engineering from SUT in 2013. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

**R. Chanklan** is currently a doctoral student with the School of Computer Engineering, SUT. She received her bachelor degree in computer engineering from SUT in 2013, master degree in Computer Engineering from SUT in 2014. Her current research of interest is data mining and artificial intelligence.

**K. Kerdprasop** is an associate professor and chair of Computer Engineering School, SUT. He received bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, MS in Computer Science from the Prince of Songkla University, in 1991, and PhD in computer science from Nova Southeastern University, U.S.A., in 1999.

**N. Kerdprasop** is an associate professor at the School of Computer Engineering, SUT. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, MS in Computer Science from the Prince of Songkla University in 1991, and PhD in computer science from Nova Southeastern University, U.S.A, in 1999.