ICIC Express Letters
Volume **13**, Number **3**, March **2019**

ICIC International ©2019 ISSN 1881-803X
pp. 255–262

PREDICTING BUSINESS CATEGORY WITH MULTI-LABEL CLASSIFICATION FROM USER-ITEM REVIEW AND BUSINESS DATA BASED ON K-MEANS

Hendry[1,2] and Rung-Ching Chen[1,*]

[1]Department of Information Management
Chaoyang University of Technology
No. 168, Jifeng East Road, Wufeng District, Taichung 41349, Taiwan
s10314905@cyut.edu.tw; *Corresponding author: crching@cyut.edu.tw

[2]Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga 50711, Indonesia
hendry@uksw.edu

Received August 2018; accepted November 2018

Abstract. *Currently, many recommendation systems propose the breakthrough of traditional single recommendation. Many items usually belong to more than one label at a time, for example, genres of music, categories of the products and emotions. One data point could be labeled more than one tag which is a problem for many classification algorithms. Clustering analysis is a primary task of data mining, which works by dividing the dataset into the partitions based on the distance of data points. Clustering is an unsupervised learning model, which is suitable to learn multi-label classification problem. The technique is commonly used in machine learning, pattern recognition, and many others. K-means is one of the simple and widely used clustering algorithms. In this paper, we propose the collaboration between business and user-item reviews to predict the multi-label classification. We implement the combination of k-means between business and user-items review. We found that the value of k equal to three will have the best multi-label classification results for business categories and business rating.*

Keywords: Multi-label classification, User-item reviews, K-means

1. **Introduction.** In recent years, information technology and big data make the data mining become an important research field. Currently, the recommendation system tries to promote the traditional single recommendation with multiple recommendations. Let data points belong to many categories. Single recommendation could not solve this challenge. In multi-label classification problems, each data point belongs to several labels at the same time; as the opposite, in traditional single-label problems, each data point only belongs to one of all possible labels [1].

Clustering is one of the important and widely used techniques to solve many problems in the fields of machine learning, pattern recognition, and data mining [2-5]. The first step of the k-means algorithm is to find the ideal k value. Several ways have been proposed to determine the k value to avoid randomly choosing by users from the input data [6,7]. The effect of this initial step is crucial, and the differences between the parameters will likely affect the results. K-means is one of the simple and widely clustering algorithms [8,9]. K-means works by choosing initial centroid from randomly selected data points. Every new data point belongs to the new cluster. The new centroid of each cluster will be recomputed by averaging the data points which are assigned into the old cluster.

Many types of research for multi-label classification have been already conducted. SVM one-against-all method [10,11] and DAGSVM [12] aim to solve multi-label classification

DOI: 10.24507/icicel.13.03.255

problems. In those experiments, all datasets are a single independent dataset. In this paper, we propose the collaboration between business and user-item review to predict the multi-label classification. We implement the combination of k-means between the review of business and user-items.

The remainders of the paper are organized as follows. Section 2 is the problem statement and the preliminary preparation. Section 3 is the proposed methodology which uses k-means for multi-label classification. Section 4 discusses the results of experiments to show the performance of the proposed methodology. Finally, we give conclusions and the future works in Section 5.

2. **Problem Statement and Preliminaries.** Clustering is one of the methodologies which divides the dataset into many clusters based on the closeness of each other and distance from those in the different clusters. For the basic clustering method, a set of n data points, $\{x_1, x_2, \ldots, x_n\}$ is clustered into k homogenous clusters by the distances of the data points' features. K-means is one of the popular clustering methods which is started by choosing random k centroids as initial centroids $\{m_1, m_2, \ldots, m_k\}$. Each data point x_i will be assigned to its closest centroid into the j^{th} cluster if the indicator function $f(j|x_i) = 1$.

$$f\left(j|x_i\right) = \begin{cases} 1, & \text{if } j = \arg \min\limits_{1 \le r \le k} ||x_i - m_r||^2 \\ 0, & \text{otherwise} \end{cases}$$

Next step, each centroid needs to be recalculated of the average of all the data points in a cluster. The candidate clusters centroid is then updated. Those steps need to be repeated until stop criterion. Stop criterions are no new reassigned values and the number of iterations is reached or all clusters centroids are converged.

The business dataset consists of the set of businesses $Bu = \{bu_1, bu_2, \ldots, bu_n\}$, where n is the total number of the businesses. User-item reviews dataset consists of the set of users $U = \{u_1, u_2, \ldots, u_m\}$, where m is the total number of users. Every user has "votes" attribute which consists of 3 properties: "Funny", "Useful", and "Cool". The other attributes are "User_id", "Stars", "date" and "Business_id". Table 1 shows the user-item reviews dataset.

Table 1. User and businesses rating relationship

votes			*User_id*	*Stars*	*date*	*Business_id*
Funny	*Useful*	*Cool*				
0	2	1	Xqd0DzHai yRqVH3W RG7hzg	5	2007-05-17	vcNAWiLM4dR 7D2nwwJ7nCA
0	2	0	H1kH6QZ V7Le4zqT RNxoZow	2	2010-03-22	vcNAWiLM4dR 7D2nwwJ7nCA
0	1	1	zvJCcrpm2 yOZrxKffw GQLA	4	2012-02-14	vcNAWiLM4dR 7D2nwwJ7nCA

In the YELP dataset, the business rating is defined by stars which are given by the calculation of the rating, review count, categories and many other attributes. In this paper, the system only considers three dimensions from the business dataset. Table 2 indicates the business dataset.

Table 2. Business dataset

Bu	*Stars*	*Review Count*	*Categories*
bu_1	4.5	1,180	$\{$list of categories$\}_1$
bu_2	2	500	$\{$list of categories$\}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
bu_n	Rating$_n$	reviewCount$_n$	$\{$list of categories$\}_n$

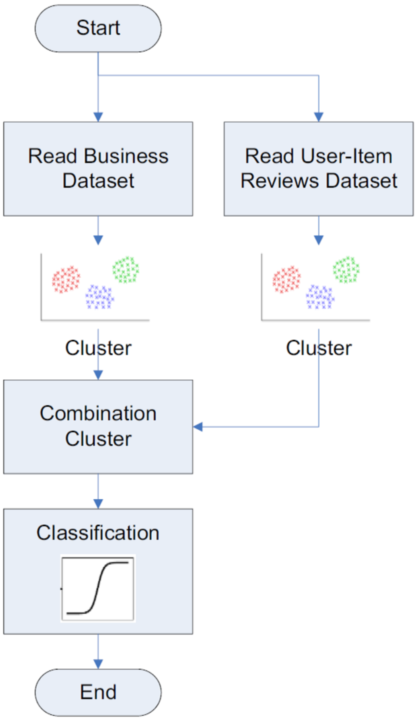


Figure 1. The system architecture

3. **Proposed Methodology.** We combine k-means clustering to do multi-label classification of YELP dataset. We collaborated with business and user-item reviews dataset. Figure 1 shows the system architecture. The system starts with data preprocessing by reading business dataset and user-itemreviews dataset. We utilize k-means clustering because it is a simple and easy algorithm. The initialization of k value is important for k-means. K-means works by averaging each data point into the nearest neighbor.

A simple cluster k-means is shown by Equation (1).

$$W(C_k) = \sum_{x_i \in C_k} ||x_i - m_i||^2, \tag{1}$$

where x_i is the data point, and m_i is the average value of data point assigned to cluster C_k. In the clustering method, we will combine the clusters from business and user-item reviews dataset as shown in Equation (2).

$$W(C_k) = \arg\max\left(\sigma\left(\sum_{x_i \in C_{1k}} ||x_i - m_{1i}||^2\right), \sigma\left(\sum_{y_i \in C_{2k}} ||y_i - m_{2i}||^2\right)\right), \tag{2}$$

where C_1 is set of cluster 1, m_1 is mean value assigned to this cluster for the business dataset, C_2 is set of cluster 2 and m_2 is mean value assigned to this cluster for user-item reviews dataset. The silhouette value is a measure of how similar an object is to its cluster compared to other clusters. The high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters [13]. $\sigma(.)$ is a silhouette function for each data point in each class; the data points will belong to one class which has a higher silhouette function value. K-means will repeat the step until all the clusters are covered or no data point can be assigned into a new cluster. The total cluster variation of the k-means could be defined in Equation (3).

$$SS(C_k) = \sum_{k=1}^{k} W(C_k), \tag{3}$$

where $SS(C_k)$ is a set of C_k. In the last step, the system will apply Support Vector Machine (SVM) classification algorithm to extracting the label from each class. We implement one vs. all SVM as a classification algorithm [14]. Given training data $(x_1, y_1), \ldots, (x_i, y_i)$, $x_i \in R^n$, $i = 1, \ldots, l$ and $y_i \in \{1, \ldots, l\}$ where y_i is the label in the class of x_i. The i^{th} SVM will be trained with all i^{th} label shown in Equation (4).

$$\text{label of } x = \arg\max_{i=1,\ldots,l} \left((\omega^i)^T \phi(x) + b^i\right), \tag{4}$$

where $(\omega^i)^T \phi(x)$ is a regularization function, and b^i is biased for every i^{th} SVM.

4. **Experiments.** The experiments used YELP dataset and selected a 746 data items of Japanese restaurant category. We used 70% data for training and 30% data for testing. We implemented a grid search to find the optimal number of clusters by frequency among the indices of the data points. The optimal number of clusters is either $k = 3$ or $k = 4$ as shown in Figure 2. The system uses $k = 3$ because it is little higher than $k = 4$. In the remainder of the experiments, we will compare the clustering results for k value equal to 2 and 3.

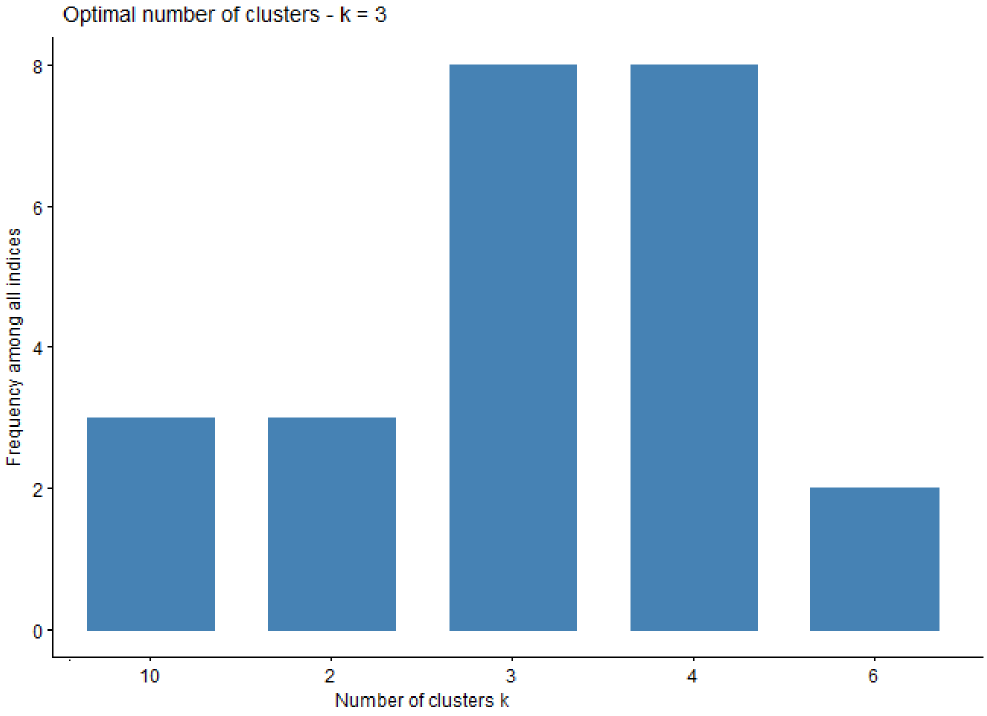
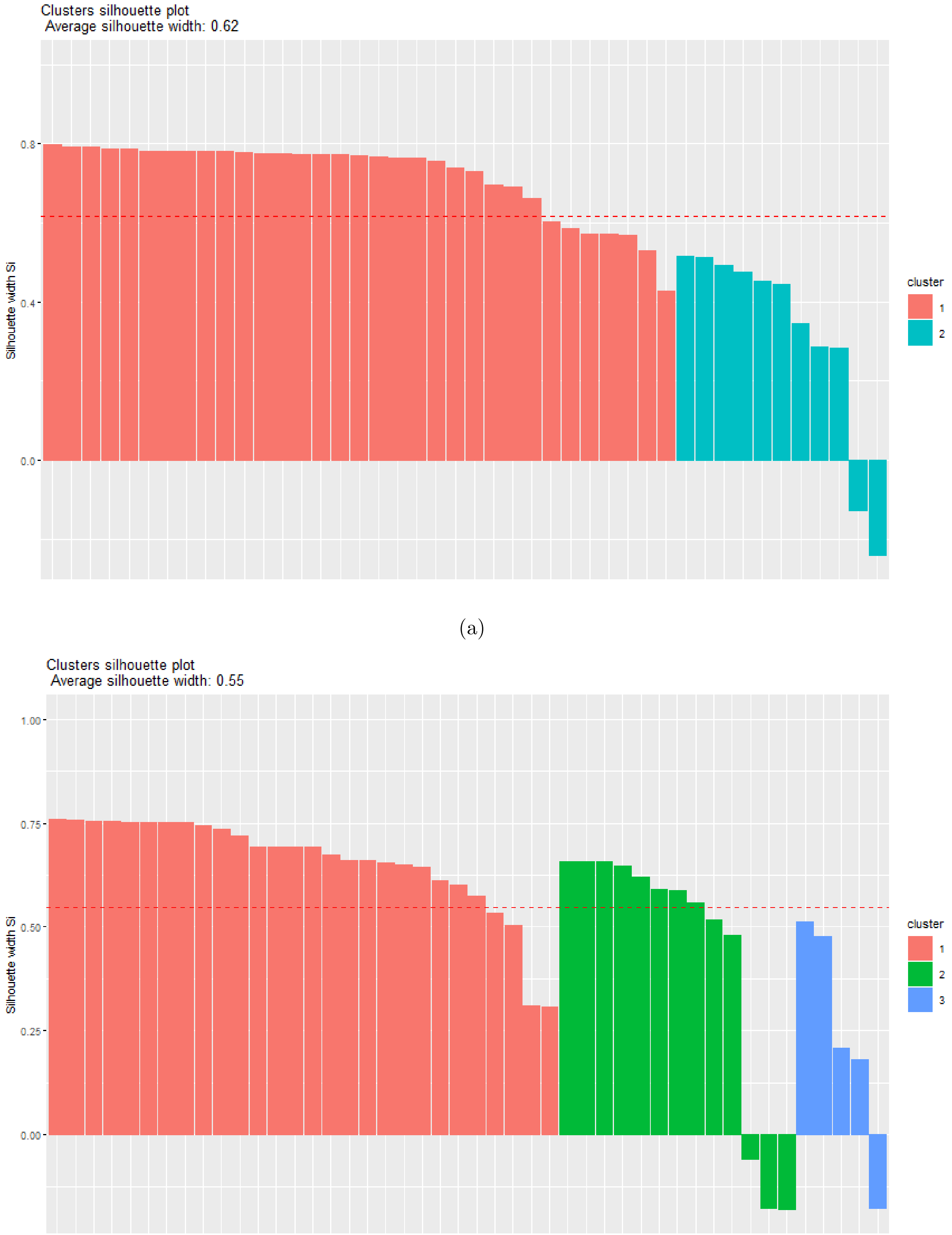


FIGURE 2. Grid search for optimal k values for the number of clusters

Figure 3 shows the single cluster silhouette value for k equal to 2 and 3. Figure 3(a) shows the silhouette of a single k-means for $k = 2$. Figure 3(b) shows the silhouette value for single k-means $k = 3$. Figure 4 shows the comparison clustering results between single and combination k-means. In single k-means, the 2nd cluster consists of fewer data points; it means maybe some data points are miss-classified and enter the other cluster. In the combination k-means, the cluster results are more balanced than single k-means.

In Figure 5, we can find the confidence data points would likely be miss-classified into the wrong cluster which is from the combination of k-means silhouette coefficients comparison between k clusters. The silhouette value for $k = 2$ is 0.79 and for $k = 3$ is 0.73.



(a)

(b)

Figure 3. Single k-means silhouette value results for different k: (a) $k = 2$, and (b) $k = 3$

Table 3. Silhoutette value distributions for cluster $k = 2$

Cluster	*Size*	*Silhouette*
1	68	0.37
2	678	0.83

Table 4. Silhoutette value distributions for cluster $k = 3$

Cluster	*Size*	*Silhouette*
1	112	0.52
2	15	0.43
3	619	0.78

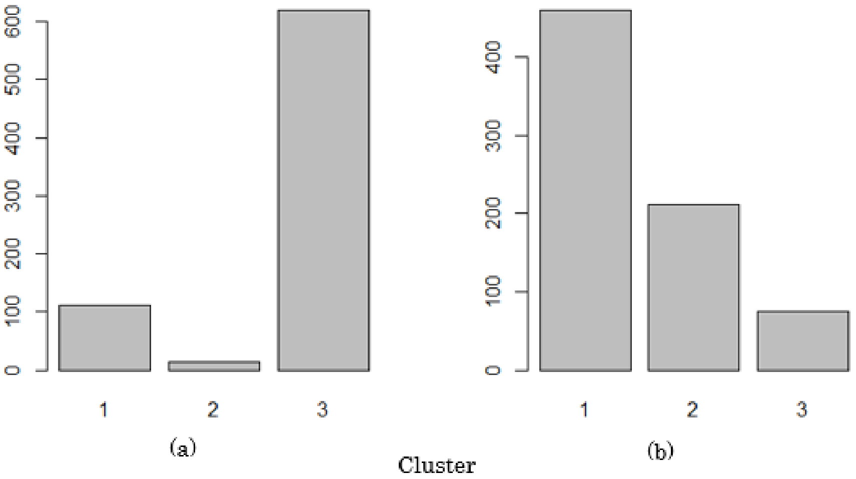


Figure 4. The comparison of clustering results between single and combination k-means: (a) single k-means, and (b) combination k-means

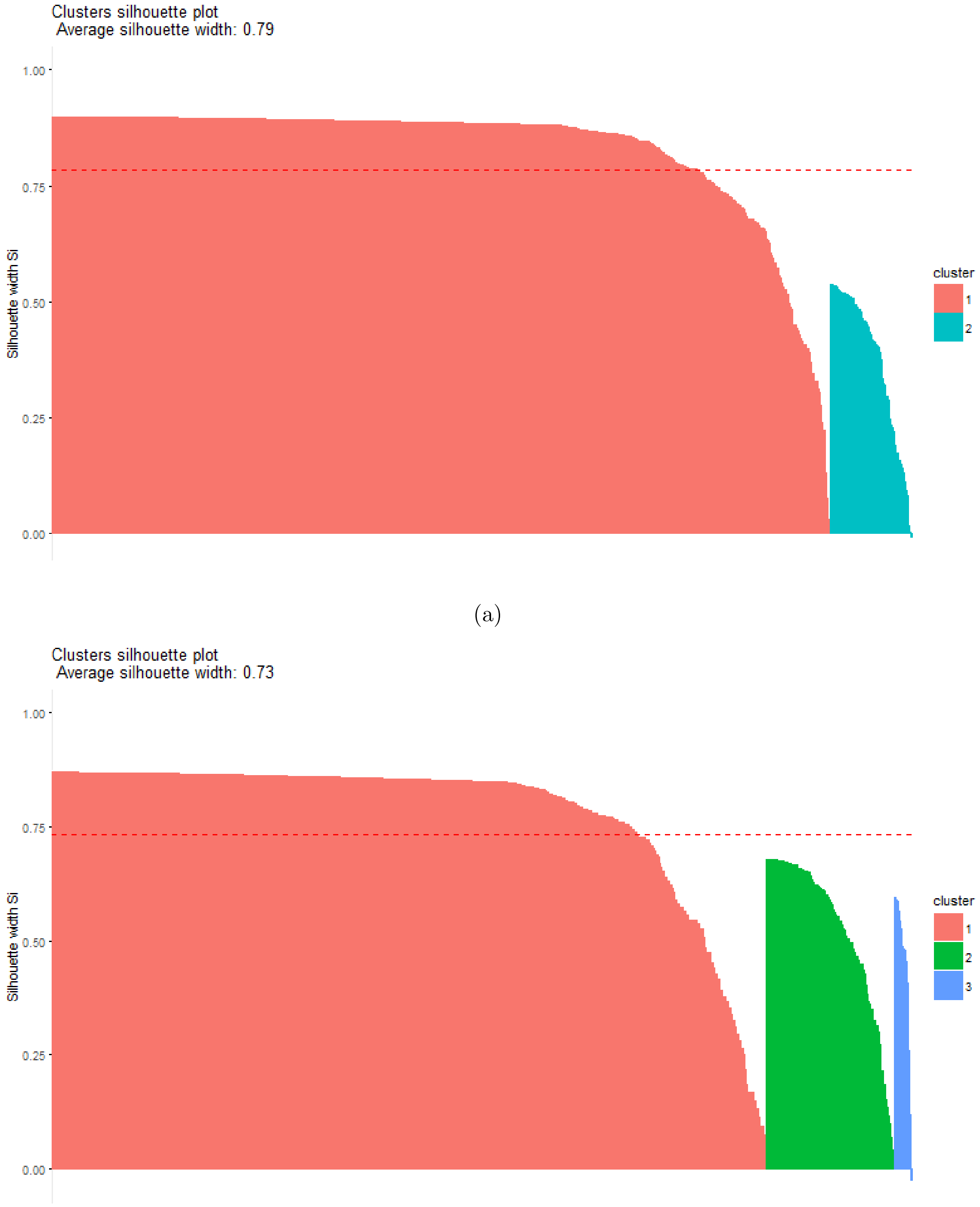
Table 5. Classification accuracy comparison for each cluster and labels

Labels	$k = 2$	$k = 3$
1	99.53	99.53
2	69.64	86.60
3	58.92	73.21

Table 3 and Table 4 are the distribution for each cluster. Although average silhouette coefficient value for $k = 2$ higher than $k = 3$, the gap between clusters are higher. Based on this result, we choose $k = 3$ as the best initial value for the model.

Table 5 shows the comparison of classification accuracy for each k value according to the number of labels classification. We find that $k = 3$ with label 2 has the highest accuracy for multi-label classification. With increasing labels number, the efficiency will be lower.

5. **Conclusions.** In this paper, we present the multi-label classification through multi k-means clustering which is utilized for business and user-item reviews. The determination of the initial k is chosen automatically by grid search to reduce the randomly chosen or trial and error process. The confidence results are compared with the silhouette coefficient and show the confidence level 0.73 with 3 clusters without overlapping class. In the future



(a)

(b)

FIGURE 5. The combination of k-means silhouette value for different k: (a) $k = 2$, and (b) $k = 3$

research, the combination of k-means techniques would be optimized using the multi-optimization algorithm instead of merging technologies. Furthermore, the initialization problems of k-means can be extended for spatial dataset where the region is considered.

Acknowledgment. This paper is supported by the Ministry of Science and Technology, Taiwan, with project number: MOST-106-2221-E-324-025 and MOST-107-2221-E-324 -018 -MY2.

REFERENCES

[1] J.-J. Zhang, M. Fang and X. Li, Clustered intrinsic label correlations for multi-label classification, *Expert Systems with Applications*, vol.81, pp.134-146, 2017.

[2] C. Bouveyron and C. Brunet-Saumard, Model-based clustering of high-dimensional data: A review, *Computational Statistics & Data Analysis*, vol.71, pp.52-78, 2014.

[3] S. Aghabozorgi, A. S. Shirkhorshidi and T. Y. Wah, Time-series clustering – A decade review, *Information Systems*, vol.53, pp.16-38, 2015.

[4] N. Ahmed, Recent review on image clustering, *IET Image Processing*, vol.9, no.11, pp.1020-1032, 2015.

[5] P. C. Besse, B. Guillouet, J. M. Loubes and F. Royer, Review and perspective for distance-based clustering of vehicle trajectories, *IEEE Trans. Intelligent Transportation Systems*, vol.17, no.11, pp.3306-3317, 2016.

[6] Y. Luo, K. Zhang, Y. Chai and Y. Xiong, Multi-parameter-setting based on data original distribution for DENCLUE optimization, *IEEE Access*, vol.6, pp.16704-16711, 2018.

[7] R. Wang, S. Lai, G. Wu, L. Xing, L. Wang and H. Ishibuchi, Multi-clustering via evolutionary multi-objective optimization, *Information Sciences*, vol.450, pp.128-140, 2018.

[8] P. Shi, X. Fan, J. Ni and G. Wang, A detection and classification approach for underwater dam cracks, *Structural Health Monitoring*, vol.15, no.5, pp.541-554, 2016.

[9] S. Khanmohammadi, N. Adibeig and S. Shanehbandy, An improved overlapping k-means clustering method for medical applications, *Expert Systems with Applications*, vol.67, pp.12-18, 2017.

[10] K. Crammer and Y. Singer, On the learnability and design of output codes for multiclass problems, *Machine Learning*, vol.47, no.2, pp.201-233, 2002.

[11] F. F. Chamasemani and Y. P. Singh, Multi-class support vector machine (SVM) classifiers – An application in hypothyroid detection and classification, *Proc. of the 6th International Conference on Bio-Inspired Computing: Theoris and Applications*, Penang, Malaysia, pp.351-356, 2011.

[12] J. C. Platt, N. Cristianini and J. Shawe-Taylor, Large margin DAGs for multiclass classification, *Proc. of the 12th International Conference on Neural Information Processing Systems*, Denver, CO, pp.547-553, 1999.

[13] R. C. de Amorim and C. Hennig, Recovering the number of clusters in data sets with noise features using feature rescalling factors, *Information Sciences*, vol.324, pp.124-145, 2015.

[14] C.-W. Hsu and C.-J. Lin, A comparison of methods for multiclass support vector machines, *IEEE Trans. Neural Networks*, vol.13, no.2, pp.415-425, 2002.