Comparison between fuzzy kernel k-medoids using radial basis function kernel and polynomial kernel function in hepatitis classification

Glori Stephani Saragih, Sri Hartini, Zuherman Rustam
Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia

Article Info

Article history:

Received Feb 20, 2020
Revised Dec 15, 2020
Accepted Jan 21, 2021

Keywords:

Classification
Fuzzy kernel k-medoids
Hepatitis
Kernel function
Polynomial kernel
Radial basis function

ABSTRACT

This paper compares the fuzzy kernel k-medoids using radial basis function (RBF) and polynomial kernel function in hepatitis classification. These two kernel functions were chosen due to their popularity in any kernel-based machine learning method for solving the classification task. The hepatitis dataset then used to evaluate the performance of both methods that were expected to provide an accurate diagnosis in patients to obtain treatment at an early phase. The data were obtained from two hospitals in Indonesia, consisting of 89 hepatitis-B and 31 hepatitis-C samples. The data were analyzed using several cases of k-fold cross-validation, and the performances were compared according to their accuracy, sensitivity, precision, F1-Score, and running time. From the experiments, it was concluded that fuzzy kernel k-medoids using RBF kernel function is better compared to polynomial kernel function with the 6% increment of accuracy, 13% enhancement of sensitivity, and 5% improvement in F1-Score. On the other side, the precision of fuzzy kernel k-medoids using polynomial kernel function is 2% higher than using the RBF kernel function. According to the results, the use of RBF or polynomial kernel function in fuzzy kernel medoids can be considered according to the primary goal of the classification.

Corresponding Author:

Glori Stephani Saragih
Department of Mathematics
Universitas Indonesia
Depok 16424, Indonesia
Email: glori.stephani@sci.ui.ac.id

1. INTRODUCTION

Hepatitis is a severe health problem and one of the leading causes of death across the globe. According to the global hepatitis report 2017 [1], approximately 257 million people were living with hepatitis B and 71 million with hepatitis C in 2015. However, in Indonesia, the prevalence of clinical hepatitis was estimated at 0.6% in 2007 [2]. These kinds of viral hepatitis tend to become chronic, thereby causing more deaths. Therefore, the prevention of viral hepatitis, as stated by Hou *et al*. [3], consists of behavior modification, passive immunoprophylaxis, and active immunization. Earlier prevention of viral hepatitis is also estimated using various machine learning techniques, which were expected to help patients take treatment in the earlier phase of the virus, thereby stopping it from being amplified [4].

Some researchers have published the use of machine learning in hepatitis classification [4-7]. In this paper, fuzzy kernel k-medoids is used to develop hepatitis classification to make it more accurate in providing a diagnosis. The kernel technique that was introduced by Vapnik [8] and later developed by Scholkopf *et al*. [9], and Christianini [10] will be used in fuzzy kernel k-medoids to overcome the possibilities of not separable linearly data set. Fuzzy kernel k-medoids have been previously used in problems related to anomaly detection [11] and multiple data detection such as breast cancer Wisconsin, diabetes, image segmentation, iris, and much more [12]. Furthermore, the machine learning method based on the kernel has previously been used in diagnosing several diseases and deliver excellent accuracy [13-17]. The kernel function is useful to avoid misclassifying the dataset with a spherical shape which is only solved by a linear function.

2. RESEARCH METHOD

2.1. Material

The hepatitis dataset, which was also used by Kurniawan and Rustam [18], was obtained from Tangerang and Mitra Keluarga Kelapa Gading Hospitals, consisting of 89 hepatitis B and 31 hepatitis C samples. Each sample is described by features such as gender, serum glutamic oxaloacetic transaminase (SGOT), serum glutamic pyruvic transaminase (SGPT), anti-HCV, HBsAg, urea, and creatinine. All of these features will be used in the process of classification.

2.2. Method

2.2.1. Fuzzy kernel k-medoids

This method is a combination of three concepts [11]. These are fuzzy k-Medoids, proposed by Krishnapuram *et al.* [19], Kernel function, which was introduced by Vapnik *et al.* [8], and fuzziness degree [20]. Given a dataset $X = \{x_1, x_2, \dots, x_n\}$ where $x_i \in \mathbb{R}^d$ for $i = 1, 2, \dots, n$. The objective function of this method is given in (1) where u_{ij} denotes the membership value of the sample x_i in the cluster j.

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij} u_{ij}^{m} K(\boldsymbol{x_i}, \boldsymbol{v_j}) \tag{1}$$

The membership value u_{ij} is updated using the formula in (2) and the medoid $\boldsymbol{v_j}$ is calculated as the formula in (3).

$$u_{ij} = \frac{\left(K(\boldsymbol{x_i},\boldsymbol{v_j})\right)^{-\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(K(\boldsymbol{x_i},\boldsymbol{v_j})\right)^{-\frac{1}{m-1}}}, 1 \le i \le n, 1 \le j \le c \tag{2}$$

$$\boldsymbol{v_j} = \boldsymbol{x_p} \text{ where } p = \arg \min_{1 \le q \le n} u_{ij}^{m} K(\boldsymbol{x_q}, \boldsymbol{x_i}) \tag{3}$$

The algorithm of fuzzy kernel k-medoids [11] is given in Figure 1.

```
Input: $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}, c, m_i, m_f, \varepsilon$, T (the maximum number of iterations allowed).
Output: $V = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots \boldsymbol{v_c}\}$, $U = [u_{ij}]$, where $1 \le i \le n, 1 \le j \le c$.
1.  Initialization: $V^0 = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots \boldsymbol{v_c}\}$
2.  $m = m_i + \frac{t}{T}(m_f - m_i)$
3.  Update membership of the data point $\boldsymbol{x_i}$ in $j^{th}$-cluster using (2).
4.  Update medoids $\boldsymbol{v_j}$ using (3).
5.  If $E = \sum_{j=1}^{c}\left(K\left(\boldsymbol{v}_j^{(t)}, \boldsymbol{v}_j^{(t-1)}\right)\right)^2 \le \varepsilon$ or $T = t$, then the iteration stops. Otherwise, $t = t + 1$ and go back to step 2;
    End.
```

Figure 1. Algorithm of fuzzy kernel k-medoids

This method utilized the RBF and polynomial kernel function. The RBF kernel mostly used because of its simplicity that has fewer hyperparameters. The number of hyperparameters used in the kernel usually influences the complexity of model selection [21]. Meanwhile, polynomial was also one of the kernel functions that commonly used mainly for the lower polynomial degree, because the infinite degree of a polynomial has the same form with the gaussian RBF kernel [22] the polynomial kernel has more hyperparameters than the RBF kernel. The formulas [23] are shown in (4-5), respectively.

RBF kernel function: $K(\boldsymbol{x_i}, \boldsymbol{v_j}) = exp\left(-\frac{\|x_i - v_j\|^2}{2\sigma^2}\right)$ (4)

Polynomial kernel function: $K(\boldsymbol{x_i}, \boldsymbol{v_j}) = (\boldsymbol{x_i} \cdot \boldsymbol{v_j} + 1)^h$ (5)

2.2.2. Research methodology

The k-fold cross-validation [24] will be used in this paper for evaluating the fuzzy kernel k-medoids algorithm. For example, when we used 3-fold cross-validation, the data is divided into three folds for each class. Therefore, we get the number of points in every fold, as shown in Table 1.

Table 1. The number of samples in every three folds of hepatitis dataset

Fold	The number of hepatitis B samples	The number of hepatitis C samples
1	30	10
2	30	10
3	29	11
Total	89	31

The k-fold cross-validation for classification tasks using fuzzy kernel k-medoids might be unfamiliar due to its utilization that commonly used for clustering or unsupervised learning [25] methods in machine learning. In this fuzzy kernel k-medoids, a fold was used to obtain the centroids of the clusters according to the algorithm in Figure 1. In contrast, the rest k−1 folds were used to evaluate the method by determining the class of every data point according to its nearest centroid. Consider the data labeled hepatitis B belongs to class 1 and the data labeled hepatitis C belongs to class 2. If the data point was nearer to the centroid of class 1, then the predicted class for this data point is hepatitis B. Meanwhile, if the data point was nearer to the centroid of class 2, then the predicted class for this data point is hepatitis C.

2.2.3. Performance measure

Accuracy, sensitivity, precision, and F1-Score were used as performance measurement. It was calculated using the (6-9) while considering the results of the confusion matrix. TP is the number of hepatitis-B samples correctly diagnosed and TN is the number of hepatitis-C samples correctly diagnosed. Meanwhile, FN is the number of hepatitis-B samples incorrectly diagnosed and FP is the number of hepatitis-C samples incorrectly diagnosed.

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \quad (6)$$

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

$$\text{F1-Score}=\frac{2 * \text{sensitivity} * \text{precision}}{\text{sensitivity} + \text{precision}} \quad (9)$$

3. RESULTS AND ANALYSIS

The performance of fuzzy kernel k-medoids is evaluated using k-fold cross-validation in which $k = 3, 5, 7, 10$. However, this research makes use of RBF and polynomial kernel function with several kernel parameters and polynomial degrees examined. The performance of fuzzy kernel k-medoids using RBF kernel function is shown in Table 2.

According to Table 2, the kernel parameter $\sigma = 0.0001$ performs excellently in every performance measurement of each cross-validation. However, the highest value of accuracy, sensitivity, precision, and F1-Score of this kernel parameter are obtained when 7-fold cross-validation is used. The performance of fuzzy kernel k-medoids using polynomial kernel function is shown in Table 3.

Table 2. The performance of fuzzy kernel k-medoids using RBF kernel function

Evaluation method	Performance measure	Kernel parameter of RBF									
		0.0001	0.001	0.05	0.1	1	5	10	50	100	1000
3-fold CV	Accuracy	78.89	77.78	77.78	77.50	74.67	72.96	73.49	72.08	71.11	70.44
	Sensitivity	98.61	97.22	96.30	95.83	90.28	86.81	87.30	84.90	83.02	82.22
	Precision	79.78	79.55	80.00	80.00	80.45	80.82	81.03	81.09	81.27	81.10
	F1-Score	88.20	87.50	87.39	87.20	85.08	83.71	84.05	82.95	82.14	81.66
	Running Time	1.27	1.03	1.13	1.13	1.50	1.13	1.11	1.08	1.06	1.08
5-fold CV	Accuracy	77.78	76.11	75.56	75.00	74.00	73.15	72.70	72.22	71.85	71.56
	Sensitivity	100.00	97.86	96.19	95.00	92.86	91.43	90.41	89.46	88.73	88.29
	Precision	77.78	77.40	77.69	77.78	77.94	77.89	77.99	78.04	78.07	78.03
	F1-Score	87.50	86.44	85.96	85.53	84.75	84.12	83.74	83.36	83.06	82.84
	Running Time	0.06	1.44	1.81	1.42	0.69	0.58	0.56	0.56	0.58	0.58
7-fold CV	Accuracy	82.14	80.95	80.56	80.65	79.76	79.37	79.08	78.87	78.70	78.57
	Sensitivity	98.57	97.14	96.19	95.71	94.29	93.57	93.06	92.68	92.38	92.14
	Precision	83.13	82.93	83.13	83.49	83.54	83.62	83.67	83.71	83.74	83.77
	F1-Score	90.20	89.47	89.18	89.18	88.59	88.31	88.12	87.97	87.85	87.76
	Running Time	2.08	1.67	1.61	1.48	0.41	0.41	0.39	0.39	0.39	0.42
10-fold CV	Accuracy	77.78	75.56	74.81	74.44	72.67	72.22	71.90	71.53	71.23	71.00
	Sensitivity	100.00	97.14	95.24	94.29	91.71	90.71	90.00	89.29	88.73	88.29
	Precision	77.78	77.27	77.52	77.65	77.35	77.44	77.50	77.52	77.53	77.54
	F1-Score	87.50	86.08	85.47	85.16	83.92	83.55	83.29	82.99	82.75	82.57
	Running Time	0.08	1.50	1.63	1.58	1.14	0.97	0.91	0.91	0.92	0.88

Table 3. The performance of fuzzy kernel k-medoids using polynomial kernel function

Evaluation method	Performance measure	Polynomial degree									
		1	2	3	4	5	6	7	8	9	10
3-fold CV	Accuracy	70.00	71.11	71.48	69.44	70.67	71.67	72.54	73.06	73.58	73.89
	Sensitivity	79.17	80.56	81.02	77.78	80.28	81.94	83.13	83.85	84.57	85.00
	Precision	82.61	82.86	82.94	82.96	82.57	82.52	82.64	82.71	82.78	82.81
	F1-Score	80.85	81.69	81.97	80.29	81.41	82.23	82.89	83.28	83.66	83.89
	Running Time	1.16	1.19	1.42	1.30	1.28	1.58	1.31	1.31	1.33	1.36
5-fold CV	Accuracy	68.89	68.89	69.26	69.17	69.56	70.56	71.27	71.81	72.22	72.56
	Sensitivity	82.86	82.86	83.33	83.21	83.43	84.52	85.31	85.89	86.35	86.71
	Precision	78.38	78.38	78.48	78.45	78.71	79.06	79.32	79.50	79.65	79.76
	F1-Score	80.56	80.56	80.83	80.76	81.00	81.70	82.20	82.58	82.86	83.09
	Running Time	0.58	0.45	1.39	0.34	1.08	1.00	1.05	1.05	1.06	1.13
7-fold CV	Accuracy	77.38	78.57	78.17	78.87	78.81	78.97	78.91	78.87	78.70	78.69
	Sensitivity	90.00	92.14	91.90	91.79	91.14	90.95	90.61	90.36	90.00	89.86
	Precision	84.00	83.77	83.55	84.26	84.62	84.89	85.06	85.19	85.26	85.35
	F1-Score	86.90	87.76	87.53	87.86	87.76	87.82	87.75	87.69	87.57	87.54
	Running Time	0.44	1.64	0.95	1.02	0.95	0.98	1.02	1.05	1.11	1.22
10-fold CV	Accuracy	68.89	68.33	68.52	69.44	70.44	71.11	71.75	72.36	72.59	72.78
	Sensitivity	84.29	82.86	82.38	83.21	84.29	85.00	85.51	86.07	86.35	86.57
	Precision	77.63	77.85	78.28	78.72	79.09	79.33	79.66	79.93	80.00	80.05
	F1-Score	80.82	80.28	80.28	80.90	81.60	82.07	82.48	82.89	83.05	83.18
	Running Time	1.30	0.84	1.11	1.14	1.22	1.08	1.58	1.66	1.72	1.66

Table 3 shows that the tenth polynomial degree almost achieves the best performance in every cross-validation. The results are more complicated in the 7-fold cross-validation because the highest value of every performance measure is obtained in a different polynomial degree. However, further analysis shows the fourth polynomial degree as the best performance following the values and the measurements. Therefore, fuzzy kernel k-medoids using RBF kernel function of σ=0.0001 and fourth polynomial kernel function are compared, as shown in Figure 2. If we analyze Tables 2-3 further in comparing each of its highest value, we can conclude that fuzzy kernel k-medoids using RBF kernel function is better compared to polynomial kernel function with the 6% increment of accuracy, 13% enhancement of sensitivity, and 5% improvement in F1-Score. On the other side, the precision of fuzzy kernel k-medoids using polynomial kernel function is 2% higher than using the RBF kernel function. Based on this figure, it is concluded that fuzzy kernel k-medoids performs better when using RBF than polynomial kernel function. The comparison shows that RBF makes fuzzy kernel k-medoids performance to become more excellent in accuracy, sensitivity, and F1-Score. On the other side, the polynomial degree makes fuzzy kernel k-medoids better in precision. The RBF kernel function performs better in these three measurements and in running time. As shown in Table 4, the fuzzy kernel k-medoids using RBF kernel function is faster in running time than the polynomial kernel function used in every evaluation method.

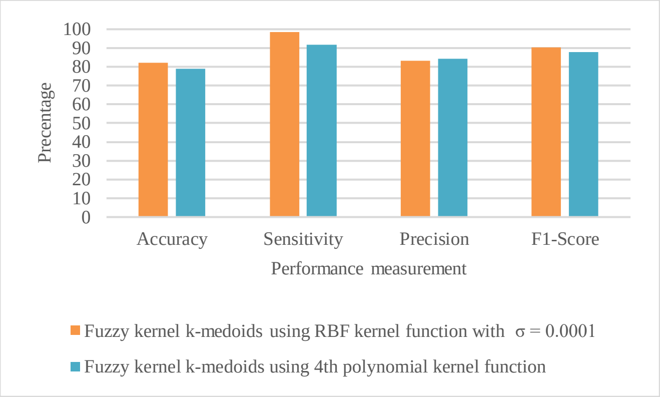

Figure 2. Comparison of fuzzy kernel k-medoids using RBF kernel function using σ=0.0001 and using the fourth polynomial kernel function

Table 4. Comparison of the best kernel function in every evaluation method

Evaluation method	Kernel function	Running time
3-fold CV	RBF kernel with σ=0.001	1.03
	1st polynomial kernel	1.16
5-fold CV	RBF with σ=0.0001	0.06
	4th polynomial kernel	0.34
7-fold CV	RBF with σ=10, 50, 100	0.39
	1st polynomial kernel	0.44
10-fold CV	RBF with σ=0.0001	0.08
	2nd polynomial kernel	0.84

4. CONCLUSION

Early detection of hepatitis is expected to help patients to obtain proper treatment, considering this disease as one of the crucial causes of death worldwide. There are several types of hepatitis; however, most found cases are hepatitis B and hepatitis C. Therefore, this paper proposed the use of the fuzzy kernel k-medoids using RBF and polynomial kernel function for the hepatitis classification. Data were obtained from two hospitals in Indonesia, consisting of 89 hepatitis-B and 31 hepatitis-C samples. According to the experiments, it is concluded that RBF using σ=0.0001 delivers better performance than the fourth polynomial kernel function in the fuzzy kernel k-medoids. Furthermore, the comparison shows that the RBF kernel makes fuzzy kernel k-medoids performance improve in accuracy, sensitivity, and F1-Score. On the other side, the polynomial degree makes fuzzy kernel k-medoids better in precision. Even though the proposed method in this paper already delivered excellent performance, the other methods with some technique to obtaining balance data can be used as future work to obtain a better, more accurate, and precise diagnosis.

ACKNOWLEDGEMENTS

The authors were grateful to the Tangerang Hospital and Mitra Keluarga Kelapa Gading Hospital for their kindness in providing the hepatitis datasets, and all the reviewers included in the improvement of this article. This research was financially supported by Universitas Indonesia, with PUTI SAINTEKES 2020 research grant scheme (ID number NKB-2380/UN2.RST/HKP.05.00/2020).

REFERENCES

[1] World Health Organization, “Global hepatitis report,” WHO, 2017.

[2] Mulyanto, “Viral hepatitis in Indonesia: Past, present, and future,” in *Euroasian Journal of Hepato-Gastroenterology*, vol. 6, no. 1, pp. 65-69, 2016, doi: 10.5005/jp-journals-10018-1171.

[3] J. Hou, Z. Liu, and F. Gu, “Epidemiology and prevention of hepatitis B virus infection,” in *International Journal of Medical Sciences*, vol. 2, no.1, pp. 50-57, 2005, doi:10.7150/ijms.2.50.

[4] K.S. Bhargav, *et al.* “Application of machine learning classification algorithms on hepatitis dataset,” in *International Journal of Applied Engineering Research*, vol. 13, no. 16, pp. 12732-12737, 2018.

[5] T. Karthikeyan, P. Thangaraju, “Analysis of classification algorithms applied to hepatitis patients,” in *International Journal of Computer Applications*, vol. 62, no.15, pp. 25-30, 2013, DOI: 10.5120/10157-5032.

[6] M. Nilashi, *et al.*, “A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique,” in *Journal of Infection and Public Health*, vol. 12, no. 1, pp. 13-20, 2019, doi:10.1016/j.jiph.2018.09.009.

[7] X. Tian, *et al.* “Using machine learning algorithms to predict hepatitis B surface antigen seroclearance,” in *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1-7, 2019, doi:10.1155/2019/6915850.

[8] V.N. Vapnik, “Statistical learning theory,” John Wiley & Sons, Hoboken, New Jersey, 1998.

[9] B. Schölkopf, A. Smola and K. Müller, “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” in *Neural Computation*, vol. 10, no. 5, pp. 1299-1319, 1 July 1998, doi: 10.1162/089976698300017467.

[10] N. Cristianini, J. S. Taylor, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods,” Cambridge University Press, Cambridge, 2000.

[11] Z. Rustam, A.S. Talita, “Fuzzy kernel k-medoids algorithm for anomaly detection problems,” in *AIP Conference Proceedings*, vol. 1862, pp. 030154, 2017, doi: 10.1063/1.4991258.

[12] Z. Rustam, A.S. Talita, “Fuzzy kernel c-means algorithm for intrusion detection systems,” in *Journal of Theoretical and Applied Information Technology*, vol. 81, no. 1, pp. 161-165, 2015.

[13] A. Wulan, M. V. Jannati, Z. Rustam and A. A. Fauzan, “Application Kernel Modified Fuzzy C-Means for gliomatosis cerebri,” *2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA)*, Banda Aceh, pp. 35-38, 2016, doi: 10.1109/ICMSA.2016.7954303.

[14] Z. Rustam, S. Hartini, “Classification of breast cancer using fast fuzzy clustering based on kernel,” in *IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 5, pp. 052067, 2019.

[15] N. Shandri, Z. Rustam, “Clustering arrhythmia multiclass using fuzzy robust kernel c-means (FRKCM),” in *2018 International Conference on Applied Information Technology and Innovation*, pp 145-148, 2018, doi: 10.1109/ICAITI.2018.8686747.

[16] Z. Rustam, A.S. Talita, “Fuzzy kernel robust clustering for anomaly based intrusion detection,” in *2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-4, doi: 10.1109/IAC.2018.8780480.*

[17] R.A. Putri, Z. Rustam, and J. Pandelaki, “Kernel based fuzzy c-means clustering for chronic sinusitis classification,” in *IOP Conference Series: Materials Science and Engineering*, vol. 546, pp. 052060, 2019, doi:10.1088/1757-899X/546/5/052060.

[18] G. Kurniawan, Z. Rustam, “Enhancement of hepatitis virus outcome predictions with application of K-means clustering,” in *AIP Conference Proceedings*, vol. 2168, no. 1, pp. 020044, 2019, doi:10.1063/1.5132471.

[19] R. Krishnapuram, A. Joshi and Liyu Yi, “A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering,” *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315)*, Seoul, South Korea, vol. 3, pp. 1281-1286, 1999, doi: 10.1109/FUZZY.1999.790086.

[20] N.B. Karayiannis, J.C. Bezdek, “An integrated approach to fuzzy learning vector quantization and fuzzy c-means clustering,” in *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 4, pp. 622–628, 1997, doi: 10.1109/91.649915.

[21] V. Apostolidis-Afentoulis, “SVM classification with linear and RBF kernels,” pp. 1-7, 2015, DOI: 10.13140/RG.2.1.3351.4083.

[22] H-Y. Huang, C-J. Lin, “Linear and kernel classification: When to use which?,” in *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 216-224, 2016, DOI:10.1137/1.9781611974348.25.

[23] L. Liu, B. Shen, and X. Wang, “Research on kernel function of support vector machine,” in *Advanced Technologies, Embedded and Multimedia for Human-centric Computing Book (Notes in Electrical Engineering)*, Eds Huang Y M, Chao H C, Deng D J, Park J, Springer, Dordrecht, 2014.

[24] Y. Jung, J. Hu, “A k-fold averaging cross-validation procedure,” in *Journal of Nonparametric Statistics*, vol. 27, no. 2, pp. 167-179, 2015, doi: 10.1080/10485252.2015.1010532.

[25] D. Greene, P. Cunningham, and R. Mayer, “Unsupervised learning and clustering,” in *Cord M., Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg,* 2008.