INT J COMPUT COMMUN, ISSN 1841-9836
9(3):370-378, June, 2014.

Comparison and Weighted Summation Type of Fuzzy Cluster Validity Indices

K.L. Zhou, S. Ding, C. Fu, S.L. Yang

Kaile Zhou*, Shuai Ding, Chao Fu, Shanlin Yang
School of Management
Hefei University of Technology
Hefei 230009, China
*Corresponding author: zhoukaile@mail.hfut.edu.cn
E-mail: dingshuai@hfut.edu.cn,
wls_fuchao@163.com, yangsl@hfut.edu.cn

Abstract: Finding the optimal cluster number and validating the partition results of a data set are difficult tasks since clustering is an unsupervised learning process. Cluster validity index (CVI) is a kind of criterion function for evaluating the clustering results and determining the optimal number of clusters. In this paper, we present an extensive comparison of ten well-known CVIs for fuzzy clustering. Then we extend traditional single CVIs by introducing the weighted method and propose a weighted summation type of CVI (WSCVI). Experiments on nine synthetic data sets and four real-world UCI data sets demonstrate that no one CVI performs better on all data sets than others. Nevertheless, the proposed WSCVI is more effective by properly setting the weights.
Keywords: fuzzy clustering, fuzzy c-means (FCM), cluster validity indices (CVIs), WSCVI.

1 Introduction

Clustering [1] is an unsupervised learning process to discover significant patterns in a given data set by partitioning a data set into groups (i.e., clusters) such that the elements assigned to the same group are as similar as possible while those in different groups are dissimilar in some sense. Clustering is an unsupervised process, the data objects in a data set are typically unlabeled and no structural knowledge about the data set is available [2]. Therefore, evaluating the quality of clustering results and determining the optimal number of clusters are difficult tasks. Also, the number of clusters is a prerequisite input parameter for many clustering algorithms [3].

Cluster validity index (CVI) is a kind of criterion function to determine the optimal number of clusters [3]. Currently, a large number of CVIs have been proposed [4, 5]. So it is necessary to evaluate and compare the performances of these CVIs. Extensive comparisons of crisp CVIs have been presented [6, 7], while few studies have focused on the performance comparison of CVIs for fuzzy clustering. Most comparison studies of fuzzy CVIs are presented when a new fuzzy CVI was proposed [8, 9], but the extent of these comparisons was limited. In this paper, we present an extensive comparative study of ten well-known fuzzy CVIs.

Previous studies on CVIs have demonstrated that there is no single CVI that can deal with any data sets and always perform better than the others [10, 11]. The idea of weighted CVIs has been mentioned in literature [12]. However, few studies have focused on the weighted summation type of CVIs for fuzzy clustering. Hence, in this paper we propose a weighted form of fuzzy clustering CVIs which is the weighted sum of ten well-known fuzzy CVIs.

The remainder of this paper is organized as follows. Section 2 reviews the fuzzy c-means clustering algorithm and ten well-known CVIs for fuzzy clustering. Then, in Section 3, we introduce the weighted summation type of fuzzy clustering CVI. Finally, experimental results are presented in Section 4. The conclusions are drawn in Section 5.

2 Fuzzy C-means and Fuzzy CVIs

2.1 Fuzzy c-means algorithm

Fuzzy c-means (FCM) algorithm [13] starts with determining the number of clusters followed by guessing the initial cluster centers. Each cluster center and corresponding membership degrees are updated iteratively by minimizing the objective functions until the termination criterion is met. The objective function of FCM is defined as:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2 \tag{1}$$

where U denotes the membership, V is the cluster center matrix, n is the number of data objects, c is the number of clusters, m is the fuzzifier in FCM, v_i is the center of cluster i, μ_{ij} is the membership degree of the jth data object x_j to v_i, d_{ij}^2 is the Euclidean distance of x_j to v_i, and $d_{ij}^2 = ||x_j - v_i||^2$.

See Ref. [13] for the iterative formulas of membership degree μ_{ij} and cluster centers v_i.

2.2 CVIs for fuzzy clustering

Proposed by Bezdek [14] in 1974, PC is the first CVI used for FCM clustering, which was defined as

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^m \tag{2}$$

The optimal cluster number c^* is obtained with the maximum value of PC.

To reduce the monotonic trend of PC index with the increase of the cluster numbers, a normalized form of PC, called NPC [15], was defined as

$$NPC = 1 - \frac{c}{c-1}(1 - PC) \tag{3}$$

The optimal number of cluster c^* is also found when NPC reach the maximum value.

The concept of entropy was introduced in PE index by Bezdek [16]. Much like PC index, PE index is defined as

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij} \log_\alpha \mu_{ij} \tag{4}$$

where α is the base of the logarithm. The optimal cluster number c^* is determined by the minimum value of PE index.

Like NPC index, NPE [17] was the normalized form of PE index to reduce the monotonic tendency of PE index and was defined as

$$NPE = \frac{n}{n-c} PE \tag{5}$$

Like PE index, the optimal cluster number is corresponding to the minimum value of NPE index.

Different from PC, NPC, PE and NPE index which only considered the membership degree elements in U, the XB index [18] consists of both the membership degree values and the information about data set itself. XB index is defined as

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2 d_{ij}^2}{n \times \min\limits_{i\neq j}||v_i - v_j||^2} \tag{6}$$

The optimal cluster number is found at the minimum value point of XB index.

Kwon [19] extended XB index and proposed a new CVI, VK, which is defined as

$$VK = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2 d_{ij}^2 + (1/c)\sum_{i=1}^{c}||v_i - \overline{v}||^2}{\min\limits_{i\neq j}||v_i - v_j||^2} \tag{7}$$

In order to determine the best clustering results and the optimal cluster number c^*, we should find the most compact and separate partition, that is, the minimum value of VK index.

Pakhira et al. [20] proposed a CVI, known as PBM index, for crisp clustering, and a corresponding form for fuzzy clustering, called PBMF index. The definition of PBMF index is

$$PBMF = (\frac{1}{c} \times \frac{E_1}{E_c} \times D_c)^2 \tag{8}$$

where $E_c = \sum_{i=1}^{c} E_i$, $E_i = \sum_{j=1}^{n} \mu_{ij} d_{ij}$, $D_c = \max\limits_{i,j=1}^{c}||v_i - v_j||^2$

The optimal cluster number is found when the maximum value of PBMF index is achieved.

Different from the ratio type of CVIs, Fukuyama and Sugeno [21] proposed a summation type of CVI called FS index. Its definition is

$$FS = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m(d_{ij}^2 - ||v_i - \overline{v}||^2) \tag{9}$$

The optimal cluster number c^* is achieved at the minimum value of FS index.

A new CVI, referred to VT index, was proposed by Tang et al. [22] based on the idea of penalty function of Kwon's index VK. VT index is defined as

$$VT = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2 d_{ij}^2 + \{1/[c(c-1)]\}\sum_{i=1}^{c}\sum_{k=1;k\neq i}^{c}||v_i - v_k||^2}{\min\limits_{i\neq k}||v_i - v_k||^2 + 1/c} \tag{10}$$

The optimal cluster number c^* is also found at the minimum value of VT index.

A CVI proposed by Bensaid et al. (SC) [23] is defined as

$$SC = \sum_{i=1}^{c}\frac{\sum_{j=1}^{n}\mu_{ij}^2 d_{ij}^2}{\sum_{j=1}^{n}\mu_{ij}\sum_{k=1}^{c}||v_i - v_k||^2} \tag{11}$$

The optimal cluster number c^* is determined by the minimum value of SC index.

3 Weighted Summation Type of Fuzzy CVIs

In order to take advantage of each fuzzy CVI and weaken its limitation, we proposed a weighted summation type of CVI (WSCVI), which is the weighted sum of the above ten fuzzy CVIs. WSCVI is defined as

$$WSCVI = \sum_{i=1}^{N} \omega_i \cdot CVI_i \tag{12}$$

where N is the number of CVIs. ω_i is the weight of index CVI_i, which represents the relative importance of the ith CVI. ω satisfies $0 \leq \omega_i \leq 1$, $\sum_{i=1}^{N} \omega_i = 1$.

CVI_i is one of the ten above CVIs for fuzzy clustering. Among them, PC, NPC and PBMF index are maximum type indices, i.e., the optimal cluster number c^* is achieved at the maximum value of these CVIs, while other seven indices are minimum type. In order to obtain c^* when WSCVI is minimum, we convert the three maximum type indices into their corresponding reciprocal types, which are presented as $PCr = 1/PC$, $NPCr = 1/NPC$, $PBMFr = 1/PBMF$. The values of different CVIs change in different range. To overcome the dominate influence of CVIs in large values, all the CVIs are normalized so that all of their values range in [0, 1].

The corresponding cluster number is optimal cluster number c^* when the values of the normalized CVIs equal to 0, and the minimum value of WSCVI is achieved.

4 Experimental Results

Nine synthetic data sets (six 2-D data sets and three 3-D data sets) and four real-world data sets were used in the experiments. The value of fuzzifier in FCM is set m=2. We suggest $\omega_1 = \omega_2 = \cdots = \omega_N = 1/N$ when there is no prior knowledge available. In the experiments, we also set some other weights to obtain the optimal cluster numbers.

The synthetic data sets are expressed as Data_d_c_n, in which d is the dimension of the data set, c is the number of clusters in the data set, and n is the total number data objects in the data set.

4.1 Data sets

Three well-known 2-D synthetic data sets, Butterfly, Example_01 and Example_02 presented in [19] and the other three 2-D synthetic data sets, Data_2_3_60, Data_2_3_70, and Data_2_4_110 [24], are shown in Figure 1 (a) to (f), respectively. Figure 2 shows the three 3-D synthetic data sets, Data_3_3_200, Data_3_3_300, and Data_3_4_320, respectively.

We also use four real-world data sets, *bupa*, *wdbc*, *iris* and *glass* data set, from UCI Machine Learning Repository [25] to test the performance of WSCVI and the ten single CVIs.

4.2 Results

The optimal cluster numbers found by the ten single CVIs and the proposed WSCVI with equal weights of each CVI are shown in Table 1. It can be seen from Table 1 that there is no one CVI that can find the optimal cluster numbers for all of the data sets. WSCVI with equal weights, $1/N$, found the correct optimal cluster numbers except Data_3_3_300, Data_3_4_320, Iris, and Glass. In order to find the optimal cluster numbers of Data_3_3_300, Data_3_4_320, Iris, and Glass using WSCVI, we adjust the weights of each CVI in WSCVI.

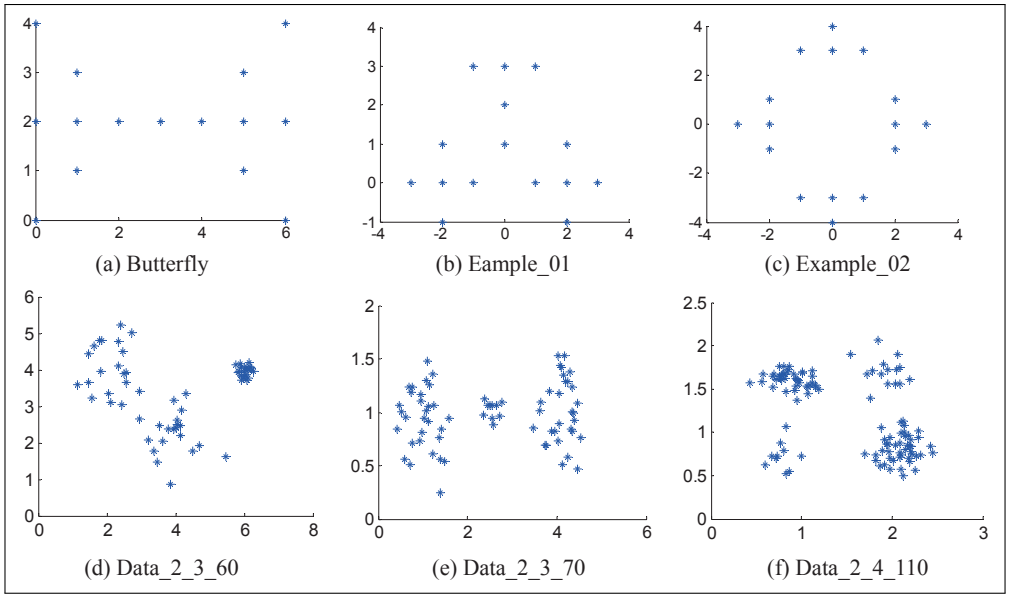

Figure 1: Six synthetic 2-D data sets

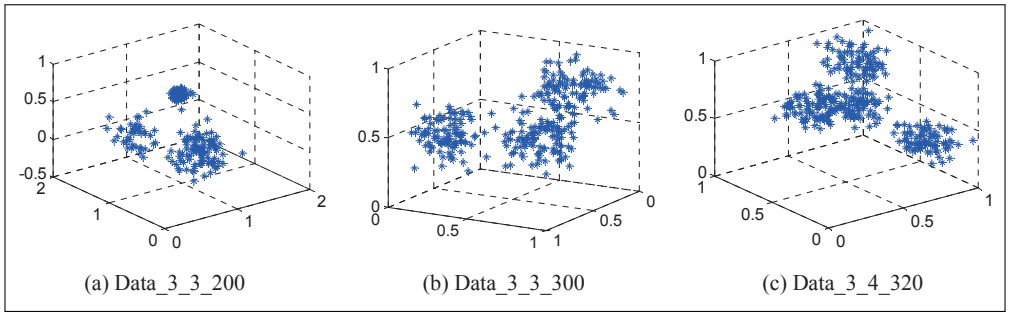

Figure 2: Three synthetic 3-D data sets

For Data_3_3_300 and Data_3_4_320, we set the weight of SC index ω_{SC}=0.5, the weights of the other CVIs are all equal to 0.056. The changes of WSCVI values with equal weights and with adjusted weights are shown in Figure 3.

As Figure 3 shows, the minimum values of WSCVI with adjusted weights are achieved at c=3 for Data_3_3_300 and c=4 for Data_3_4_320. Therefore, the optimal cluster number c^*= 3 for Data_3_3_300 and c^*= 4 of Data_3_4_320 are both found.

For iris data set, the optimal cluster number c^*= 2 considering the geometric structure is found by WSCVI when all the weights are equal to 0.1. Now we consider an adjusted weights case, in which the weight of PBMF index ω_{PBMF}=0.7, and the weights of other indices are equal to 0.025. The changes of WSCVI values with equal weights and adjusted weights on iris data are shown in Figure 4 (a).

As shown in Figure 4 (a), the optimal cluster number c^*= 3 for iris can be found when one CVI dominate other CVIs. Since six classes in glass data set are heavily overlapped, it is difficult to find six clusters. The changes of WSCVI values with equal weights and adjusted weights for glass data are shown in Figure 4 (b).

Table 1: Optimal cluster numbers preferred by each CVI

	c^*	PC	NPC	PE	NPE	XB	VK	PBMF	FS	VT	SC	WSCVI
Butterfly	2	2	2	2	2	2	2	2	2	2	2	2
Example_01	3	3	3	3	2	3	3	2	3	3	3	3
Example_02	4	4	4	4	2	4	4	2	4	4	4	4
Data_2_3_60	3	3	3	3	3	3	3	2	7	3	3	3
Data_2_3_70	3	2	3	2	2	2	2	2	4	2	4	3
Data_2_4_110	4	2	4	2	2	2	2	2	4	2	10	4
Data_3_3_200	3	3	3	2	2	3	3	2	10	3	3	3
Data_3_3_300	3	2	2	2	2	2	2	2	15	2	3	2
Data_3_4_320	4	3	4	2	2	3	3	2	5	3	4	3
Bupa	2	2	2	2	2	2	2	3	4	2	2	2
Wdbc	2	2	2	2	2	2	2	5	4	2	2	2
Iris	3	2	2	2	2	2	2	3	5	2	2	2
Glass	6	2	2	2	2	2	2	6	6	2	5	2

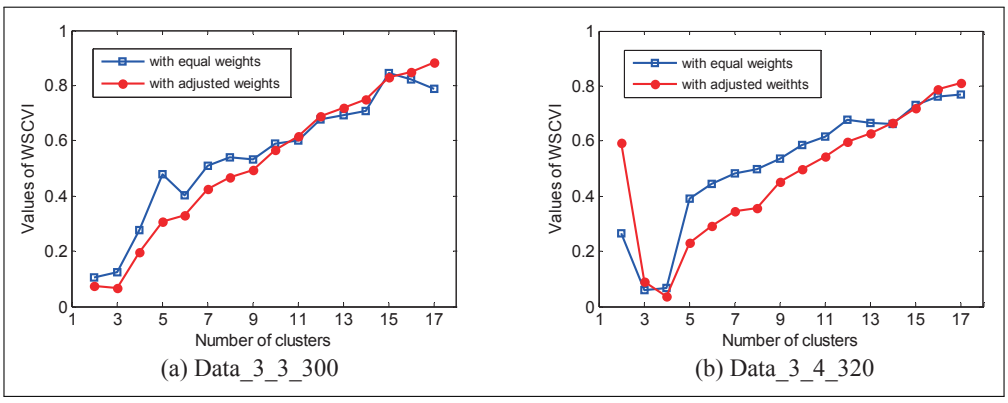

Figure 3: Values of WSCVI

From Figure 4 (b), the minimum value of WSCVI with adjusted weights achieved at c=4, and its value is the second smallest when c=6. Also, there is a large increase when c is greater than 6. Therefore, WSCVI index with adjusted weights offers the information that c^*=6 is a good cluster number estimate for glass data set.

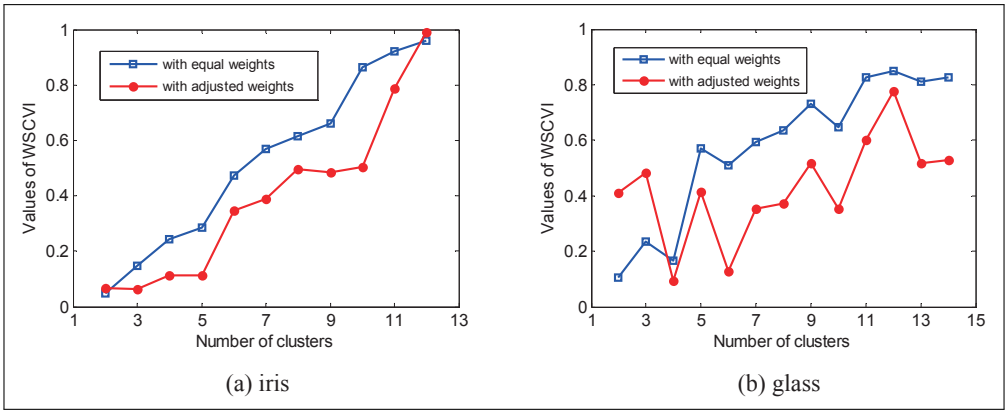

Figure 4: Values of WSCVI

With equal weights for nine data sets and adjusted weights for four data sets of each CVI, WSCVI finally find all the optimal cluster numbers for the above thirteen data sets.

5 Conclusion

We investigate ten well-known CVIs for fuzzy clustering and present an extensive comparison of the ten single CVIs and the proposed WSCVI on nine synthetic data sets and four real-world data sets. Experimental results demonstrate that most single fuzzy CVIs are effective in finding optimal cluster numbers for data sets which are low-dimensional and well-separated. But some CVIs fail to find the optimal cluster numbers for some high-dimensional or heavily overlapped data sets. The experimental results indicate that, by properly setting the weights of each CVI, the proposed WSCVI is more effective in finding the optimal cluster numbers than single CVI.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (71131002, 71201042).

Bibliography

[1] A.K. Jain, M.N. Murty, P.J. Flynn (1999). Data Clustering: A Review, *ACM Computer Surveys*, 31(3):264-323.

[2] P.A. Devijver, J. Kittler (1982). Pattern Recognition: A Statistical Approach, *Prentice-Hall*, London.

[3] F. Hoppner, F. Klawon, R. Kruse, T. Runkler (1999). Fuzzy Cluster Analysis: Methods for Classifications Data Analysis and Image Recognition, *Wiley*, New York.

[4] M. Kim, R.S. Ramakrishna (2005). New Indices for Cluster Validity Assessment, *Pattern Recognition Letters*, 26 (15):2353-2363.

[5] W. Wang, Y. Zhang (2007). On Fuzzy Cluster Validity Indices, *Fuzzy Sets and Systems*, 158(19):2095-2117.

[6] E. Dimitriadou, S. Dolnicar, A. Weingessel (2002). An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets, *Psychometrika*, 67(1):137-159.

[7] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. P¨Śrez, I. Perona (2013). An Extensive Comparative Study of Cluster Validity Indices, *Pattern Recognition*, 46(1):243-256.

[8] K.L. Wu, M.S. Yang (2005). A Cluster Validity Index for Fuzzy Clustering, *Pattern Recognition Letters*, 26 (9):1275-1291.

[9] H. Le Capitaine, C. Frelicot (2011). A Cluster-validity Index Combining an Overlap Measure and a Separation Measure based on Fuzzy-aggregation Operators, *IEEE Transactions on Fuzzy Systems* , 19(3):580-588.

[10] U. Maulik, S. Bandyopadhyay (2002). Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1650-1654.

[11] K.R. Zalik (2010). Cluster Validity Index for Estimation of Fuzzy Clusters of Different Sizes and Densities, *Pattern Recognition*, 43(10):3374-3390.

[12] W. Sheng, S. Swift, L. Zhang, X. Liu (2005). A Weighted Sum Validity Function for Clustering with a Hybrid Niching Genetic Algorithm, *IEEE Transactions on Systems, Man, and Cybernetics - Part B, Cybernetics*, 35(6):1156-1167.

[13] J.C. Bezdek, R. Ehrlish, W. Full (1984). FCM: The Fuzzy C-means Clustering Algorithm, *Computers & Geosciences*, 10(2-3):191-203.

[14] J.C. Bezdek (1974). Numerical Taxonomy with Fuzzy Sets, *Journal of Mathematical Biology*, 7(1):57-71.

[15] M. Roubens (1978). Pattern Classification Problems and Fuzzy Sets, *Fuzzy Sets and Systems*, 1(4):239-253.

[16] J.C. Bezdek (1974). Cluster Validity with Fuzzy Sets, *Journal of Cybernetics*, 3(3):58-72.

[17] J.C. Dunn (1977). Fuzzy Automata and Decision Processes, *Elsevier*, New York.

[18] X.L. Xie, G. Beni (1991). A Validity Measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841-847.

[19] S.H. Kwon (1998). Cluster Validity Index for Fuzzy Clustering, *Electronics Letters*, 34(22):2176-2177.

[20] M.K. Pakhira, S. Bandyopadhyay, U. Maulik (2004). Validity Index for Crisp and Fuzzy Clusters, *Pattern Recognition*, 37(3):487-501.

[21] Y. Fukuyama, M. Sugeno (1989). A New Method of Choosing the Number of Cluster for the Fuzzy C-means Method, *Proceedings of the 5th Fuzzy Systems Symposium*, 247-250.

[22] Y.G. Tang, F.C. Sun, Z.Q. Sun (2005). Improved Validation Index for Fuzzy Clustering, *American Control Conference*, 1120-1125.

[23] A.M. Bensaid, L.O. Hall, J.C. Bezdek, L.P. Clarke, M.L. Silbiger, J.A. Arrington, R.F. Murtagh (1996). Validity-guided (Re) Clustering with Applications to Image Segmentation, *IEEE Transactions on Fuzzy Systems*, 4(2):112-123.

[24] K.L. Zhou, S.L. Yang (2013). A Fuzzy Cluster Validity Index in Consideration of Different Size and Density of Data Set, *Journal of the China Society for Scientific and Technical Information*, 32(3):306-313.

[25] A. Asuncion, D.J. Newman (2007). UCI Machine Learning Repository, *University of California, School of Information and Computer Science*, Irvine, CA, http://www.ics.uci.edu/mlearn/MLRepositor-y.html.