Discovery of Arbitrary-Shapes Clusters Using DENCLUE Algorithm

Mariam Khader[1] and Ghazi Al-Naymat[2,1]
[1]Departmentof Computer Science, Princess Sumaya University for Technology, Jordan
[2]Department of IT, Ajman University, UAE

Abstract: *One of the main requirements in clustering spatial datasets is the discovery of clusters with arbitrary-shapes. Density-based algorithms satisfy this requirement by forming clusters as dense regions in the space that are separated by sparser regions. DENCLUE is a density-based algorithm that generates a compact mathematical form of arbitrary-shapes clusters. Although DENCLUE has proved its efficiency, it cannot handle large datasets since it requires large computation complexity. Several attempts were proposed to improve the performance of DENCLUE algorithm, including DENCLUE 2. In this study, an empirical evaluation is conducted to highlight the differences between the first DENCLUE variant which uses the Hill-Climbing search method and DENCLUE 2 variant, which uses the fast Hill-Climbing method. The study aims to provide a base for further enhancements on both algorithms. The evaluation results indicate that DENCLUE 2 is faster than DENCLUE 1. However, the first DECNLUE variant outperforms the second variant in discovering arbitrary-shapes clusters.*

Keywords: *Clustering, DENCLUE, Density Clustering, Hill-Climbing.*

Received January 24, 2020; accepted June 9, 2020
https://doi.org/10.34028/iajit/17/4A/7

1. Discovery of Arbitrary-Shapes Clusters

Cluster analysis is important to understand the distribution of data and has been applied in many applications [4, 9]. Density-based clustering algorithms have gained substantial attention in both theory and practice, due to their efficiency in discovering clusters with arbitrary shapes [14, 17]. Density-based clustering algorithms explore the data space at high scales of granularity. Generally, density-based algorithms explore the data space at a rationally high scale of granularity and then a post-processing step is applied to merge the dense regions of the explored data space to form arbitrary shapes. The advantage of such approach is that it can reconstruct the true shape of the data distribution [1].

The density at a specific point in the space is estimated, based on local connectivity (number of points in a predetermined extent of the point locality) or based on a density function. DBSCAN, which is a local connectivity based algorithms, estimate the density by determining the number of points in a fixed-radius neighbourhood. A point *x* is surrounded by a dense region if its ε neighbourhood contains at least δ number of points, including *x* itself [10]. The DENCLUE algorithm estimates the density using a density function, the "Kernel Density Estimation" (KDE). The density function at a point x is estimated as the sum of the influence functions of all data points at the point x. After that, x is assigned to an attractor point, x*, which is the "local maximum" of the density function at the point x. If the density of x*exceeds the density threshold, ξ, then x is considered within a dense region [1].

DENCLUE was first proposed by [7], in which the Hill-Climbing method was used to find the attractors (local maximum) of the density functions. The DENCLUE has good features in comparison with other density-based algorithms.

1. It uses a firm mathematical base for density estimation.
2. Achieves high performance in datasets with large amounts of noise.
3. It employs a compact mathematical form of arbitrary-shapes clusters in high-dimensional datasets.
4. Outperforms other algorithms in terms of execution time.

Different enhancements have been applied on DENCLUE algorithm [10]. Hinneburg and Gabriel [6] proposed an updated Hill-Climbing method, named "Fast Hill-Climbing", which requires less number of iterations to find the local maximum. Therefore, the fast Hill-Climbing method is faster than the gradient-based Hill-Climbing method in finding density-attractors.

This study aims to discuss the difference between the two algorithms proposed in [6, 7]:

- Hinneburg and Keim [7], reports the first variant of the DENCLUE algorithms, which applies the gradient-based Hill-Climbing method to find density-attractors.

- In [6], a direct update rule of the Hill-Climbing method was proposed. In passing, notice this updated rule is essentially the Mean-Shift algorithm proposed by [5], as testified by [19].
- If multiple modes occur in a cluster, the Mean-Shift may fail to detect the correct structure of the cluster [16]. Consequently, DENCLUE 1 outperforms DENCLUE 2 in discovering such types of clusters as shown in the experiments.
- DENCLUE 2 outperforms DENCLUE 1 in terms of execution speed, due to the updated Hill-Climbing method and K-means sampling.

The study aims to provide a base for further enhancements on both algorithms. In addition, it will help to determine which variant can be used for a particular dataset. To evaluate the result of the clustering of both algorithms, we have used the Adjusted Rand Index.

The structure of this study is as follows: Section 2 provides an overview of the DENCLUE algorithm. The steps of density clustering of the DENCLUE 1 algorithm are discussed in section 2.1 and the steps of density clustering of Denclue 2 are discussed in section 2.2. Section 3 discusses the behaviour of the DENCLUE 2 algorithm and section 4 presents the empirical evaluation of the algorithms. Finally, section5 concludes this study.

2. DENCLUE Algorithms

This section summarizes the both variants of the DENCLUE algorithm, DENCLUE 1 and DENCLUE 2.

2.1. DENCLUE 1

The DENCLUE algorithm estimates the density of a data point as the sum of the influence of all other data points in the dataset. The influence of a data point is modelled via a kernel function, such as the Gaussian kernel. The sum of all kernels (using suitable normalization) provides an estimation of the probability at any point *x*,

$$\hat{p}(x) = \frac{1}{(nh^d)} \sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \quad (1)$$

The gradient of $\hat{p}$is:

$$\frac{1}{h^{d+2}N} \sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) . (x - x_i) \quad (2)$$

Formally, a density-attractor *x**of a given influence function is the local maximum of the density-function p. Any point, *x*, is considered density-attracted to the density-attractor x*iff∃ K ∈ N:d(x^k,x*) ≤ε and satisfies the following:

$$x^0 = x, x^{1+1} = x^i + \delta . \frac{\nabla \hat{p}(x^i)}{\|\nabla \hat{p}(x^i)\|} \quad (3)$$

DENCLUE uses the parameter δ to control the speed of convergence, $\nabla \hat{p}(x^i)$is the gradient of the density function $\hat{p}(x^i)$, and $\|\nabla \hat{p}(x^i)\|$ is the euclidean norm of the gradient density function. The Hill-Climbing search stops at the step k, such that k>0 and satisfies the condition $\hat{p}(x^{k+1}) < \hat{p}(x^k)$.After that, it assigns the point x to the density-attractor x*= x^k. The parameter ξ is used to determine the minimum density threshold. A point x is considered noise or outlier if it converges in the Hill-Climbing search to a local maximum which density is less than the minimum density threshold, p(x*) <ξ.

A "center-defined" and "arbitrary-shape" clusters are defined as follows:

- A center defined cluster, (wrtσ and ξ) is a subset in which every point in this subset has isattracted tox*and has density ≥ ξ.
- An arbitrary-shape cluster (wrt σ and ξ) is a set of density-attractors X, such that: For every two attractors x_1*, x_2* ∈ X, there exist a path P from x_1* to x_2*, with all p ∈ P, $\hat{p}(p) \geq \xi$.

The number of discovered clusters via such approach varies depends on σ [7].

2.2. DENCLUE 2

The DENCLUE 2 algorithm uses an updated Hill-Climbing method for finding attractors. The updated method modifies the step size without any extra cost, finds the local maximum exactly and requires less number of iterations. Instead of using the "gradient-based" Hill-Climbing, the first derivative of $p(x)$is set to zero and the equation is solved for *x*, such that:

$$x = \frac{\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right).x_i}{\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)} \quad (4)$$

The updated Hill-Climbing does not require the step size δ and the assignment of points to clusters is performed in a different way. The search starts at a point x_i ∈X and keeps iterating until the density of attractor does not change much, such that, $\left[p\left(x_i^l\right) - p\left(x_i^{l-1}\right)\right]/p\left(x_i^l\right) \leq \varepsilon$. A new measure is calculated, which is the size of the last two steps to the attractor x*, denoted bys_t. This measure is calculated as$s_t = \sum_{i=1}^{k} \left\|x_i^{l-i+1} - x_i^{l-1}\right\|$.

Two points x_1 and x_2 are in the same cluster if their attractors,$x_1^*$and x_2^*fall within a particular distance of each other, $|x_1^* - x_2^*| \leq s_{t1} + s_{t1}$.

In their research, [6] proved the convergence of the updated hill climbing using Gaussian Kernel by casting the density function maximization as a special case of the EM algorithm. In order to enhance the performance of the algorithm, an acceleration based on sampling, using the K-means algorithm was used.

3. The Behaviour Analysis of the DENCLUE 2

The Hill-Climbing is a critical step in the DENCLUE algorithm, and selecting an appropriate value for the step size affects the discovery of density-attractors. If the step size is too large, the density-attractor may be missed. However, selecting a small value for the step size, too many steps will be needed to find the density-attractor [12]. To overcome this issue, Hinneburg and Gabriel [6] proposed the updated Hill-climbing method as an alternative to the "gradient-based" Hill-climbing method.

This proposed method, the fast Hill-climbing method, is essentially the Mean-Shift method, which was proposed by [5], as testified by [15, 19]. The Mean-Shift method is deterministic (i.e., it has no step size) and nonparametric (does not require to determine the number of the clusters in advance) [2, 11].

The Mean-Shift may fail to capture the correct structure of the cluster if the cluster contains multiple modes. For instance, a Spiral dataset contains continuous dense-regions in each cluster, which may result in detecting multiple modes by the Mean-Shift [16]. As a result, the algorithm may fail to discover the true structure of clusters with arbitrary shapes.

We have conducted experiments to trace the movement of the attractors found by DENCLUE 1 and DENCLUE 2. Two points were selected from the Spiral dataset for the experiments. The first pointis the point 90 (22.9, 16.9) and the second is the point 272(25.75, 13.7). The points were selected randomly from two different clusters. Figure 1 shows the attractors movement of two points in the Spiral dataset using DENCLUE 1. Figure 2 shows the attractors movement of the same two points in the Spiral dataset using DENCLUE 2 (without using sampling). Since DENCLUE 2 starts by a large step (and we cannot control the steps size), it has missed the correct density-attractor of the second point and assigned it to the same cluster of the first point.

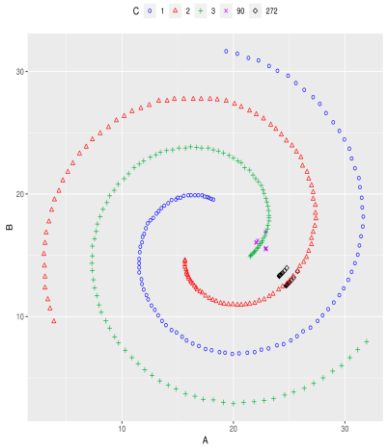

Figure 1. Attractor movement using hill-climbing (DENCLUE 1).

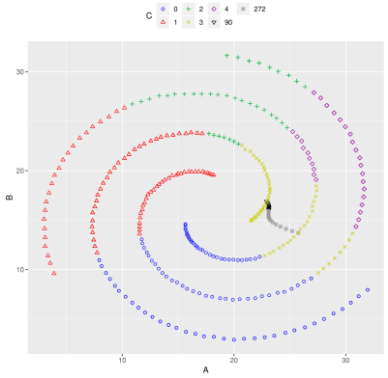

Figure 2. Attractor movement using mean-shift (DENCLUE 2).

4. Empirical Evaluation

To evaluate the DENCLUE performance, different empirical evaluations have been conducted on the performance of DENCLUE 1 and DENCLUE 2 (with sampling). Both algorithms used in the experiments are implemented using Java. The experiments were run on a machine with 7-core 2.6-GHz CPU, 8 GB of RAM and a 1-TB hard disk.

4.1. Clustering Quality

The first set of experiments were conducted to evaluate the quality of the clustering using the Adjusted Rand Index (ARI) [8]. The used datasets are the three 2D datasets: 3-Spiral dataset in Figure 3, which was used in [3] and two datasets generated using "Scikit Learn" library*https://scikit-learn.org*, the Noisy Two-Circles Dataset, Figure 4 and the Noisy Two-Moons Dataset, Figure 5.

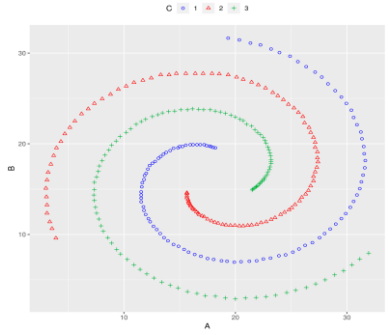

Figure 3. Spiral dataset.

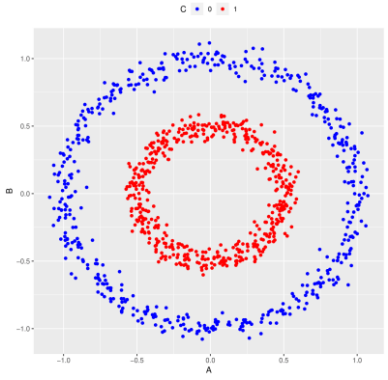

Figure 4. Noisy two-circles dataset.

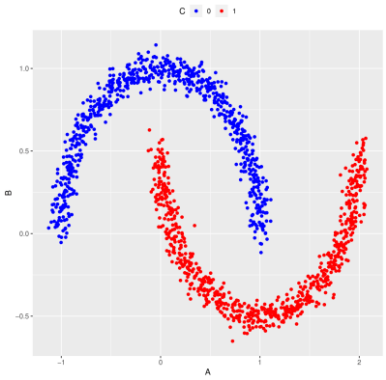

Figure 5. Noisy two-moons dataset.

Besides the mentioned datasets, five d-dimensional datasets were used. All used datasets are described in Table 1, including the number of objects (instances), the number of dimensions, and the number of clusters of each dataset.

Table 1. Description of the datasets.

Dataset	Number of Instances	Number of Dimensions	Number of Clusters
Noisy Two-Circles	1024	2	2
Noisy Two-Moons	1500	2	2
3-Spiral	312	2	3
R15	600	2	17
Compound	399	2	6
Iris	150	4	3
Seeds	210	7	3
Segment	2310	19	7
WDBC	569	30	2
Libras	360	90	15

The results of clustering were measured via the ARI measure and shown in Table 2. As a cluster validation measure, the ARI has proved its success among other measures [13, 18]. It is used to compare the clustering results of an algorithm against external criteria and it lies between zero and one. If the ARI value is near 1, this means that the two partitions have achieved better ARI value. The ARI is computed based on the following Equation:

$$\text{ARI} = \frac{\text{Index} - \text{Expected Index}}{\text{MaxIndex} - \text{ExpectedIndex}} \quad (5)$$

The results in Table 2 indicates that DENCLUE 1 outperforms DENCLUE 2, specifically, in datasets that contain clusters with multiple modes, such as Noisy Two-Circles, Noisy Two-Moons, and 3-Spiral datasets.

Carreira-Perpinan in [2] proves that the Gaussian Mean-Shift algorithm is a particular instance of the EM algorithm. That, if the kernel, used by Mean-Shift, is Gaussian, then the Mean-Shift algorithm is an Expectation-Maximization (EM) algorithm. This explains the behaviour of DENCLUE 2. Since the EM algorithm produces globular-shapes clusters, it cannot deal well with datasets that contain arbitrary-shapes clusters [1].

This can be seen in Figures 6-11, which are the visual representation of the clustering results of both DENCLUE 1 and DENCLUE 2 (without sampling) on the Noisy Two-Circles, the Noisy Two-Moons, and the 3-Spiral datasets.

Table 2. The Adjusted Rand Index value of DENCLUE 1 and DENCLUE 2.

Algorithm	Dataset	ARI value	Parameters
DENCLUE 1	Noisy Two-Circles	1.0000	$\sigma = 0.15$, $\xi = 3$
	Noisy Two-Moons	1.0000	$\sigma = 0.2$, $\xi = 3$
	3-Spiral	1.0000	$\sigma = 2.5$, $\xi = 3$
	R15	0.6749	$\sigma = 0.8$, $\xi = 3$
	Compound	0.7862	$\sigma = 1.2$, $\xi = 3$
	Iris	0.6988	$\sigma = 23$, $\xi = 3$
	Seeds	0.7810	$\sigma = 0.7$, $\xi = 3$
	Segment	0.9838	$\sigma = 1.5$, $\xi = 3$
	WDBC	0.6377	$\sigma = 20$, $\xi = 10$
	Libras	0.9251	$\sigma = .3$, $\xi = 3$
DENCLUE 2	Noisy Two-Circles	-0.0009	$\sigma = .39$
	Noisy Two-Moons	0.3406	$\sigma = 0.39$
	3-Spiral	0.0267	$\sigma = 2.5$
	R15	0.2550	$\sigma = 0.5$
	Compound	0.5681	$\sigma = 0.9$
	Iris	0.4338	$\sigma = 85$
	Seeds	0.6913	$\sigma = 0.7$
	Segment	0.194	$\sigma = 2.4$
	WDBC	0.3007	$\sigma = 18$
	Libras	0.9787	$\sigma = .3$

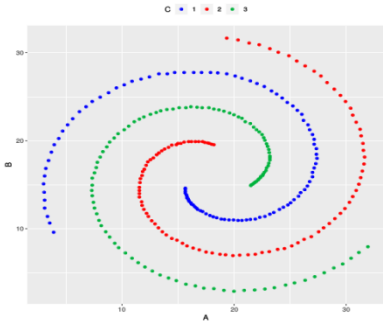

Figure 6. DENCLUE 1-spiral dataset.

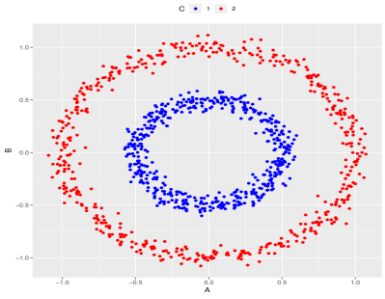

Figure 7. DENCLUE 1-circles dataset.

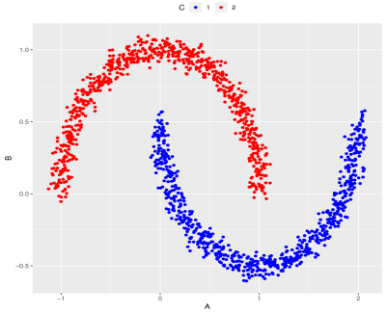

Figure 8. DENCLUE 1-moons dataset.

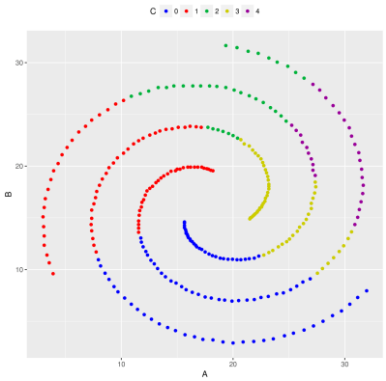

Figure 9. DENCLUE 2-spiral dataset.

All conducted experiments indicate that DENCLUE 1 can detect clusters with arbitrary-shapes more accurately than DENCLUE 2.

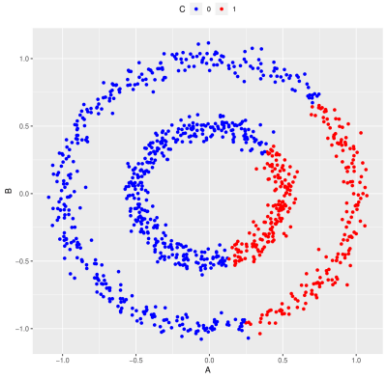

Figure 10. DENCLUE 2-circles dataset.

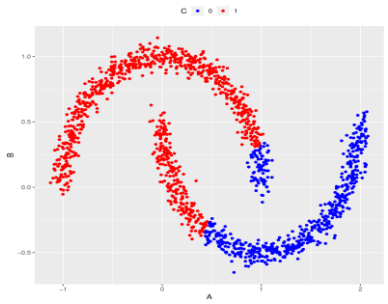

Figure 11. DENCLUE 2-moons dataset.

4.2. Clustering Performance

The second experiments set was conducted to assess the performance of both algorithms. The results of their running time are shown in Table 2. DENCLUE 2 (without sampling) was faster than DENCLUE 1 in all experiments because of using the fast Hill-Climbing method, which reduces the number of steps needed to find the density attractor. Because of using sampling, DENCLUE 2 can reduce the running time almost to the half. Hence, for datasets with spherical-shapes clusters, DENCLUE 2 is preferable.

5. Conclusions

Due to the importance of discovering arbitrary-shapes clusters in spatial datasets, it has attracted a lot of attention. DENCLUE algorithm provides an efficient solution to discover arbitrary-shapes clusters via a mathematical basis. The algorithm DENCLUE 2 was proposed to enhance the performance of DENCLUE 1 based on the Hill-Climbing method using a fast Hill-Climbing method. Although this variant has improved the performance of DENCLUE, it may fail to capture the correct structure of clusters with arbitrary-shapes. The results of the evaluation indicate that DENCLUE 2 performance outperforms DENCLUE 1 in terms of execution time. However, the DENCLUE 1 outperforms the DENCLUE 2 in discovering arbitrary-shapes clusters.

Table 3. The Running Time (in seconds) of DENCLUE 1 and DENCLUE 2 Algorithms.

Algorithm	Dataset	Running Time
DENCLUE 1	Noisy Two-Circles	2.943
	Noisy Two-Moons	8.626
	3-Spiral	0.224
	R15	1.200
	Compound	0.102
	Iris	0.310
	Seeds	0.172
	Segment	0.387
	WDBC	62.376
	Libras	0.316
DENCLUE 2	Noisy Two-Circles	0.362
	Noisy Two-Moons	0.152
	3-Spiral	0.269
	R15	0.069
	Compound	0.840
	Iris	0.113
	Seeds	0.170
	Segment	1.240
	WDBC	1.821
	Libras	1.806
DENCLUE 2 With Sampling	Noisy Two-Circles	0.316
	Noisy Two-Moons	0.417
	3-Spiral	0.089
	R15	1.609
	Compound	0.165
	Iris	0.387
	Seeds	5.159
	Segment	0.224
	WDBC	1.200
	Libras	0.102

References

[1] Aggarwal C. and Reddy C., *Data Clustering: Algorithms and Applications*, CRC Press, 2013.

[2] Carreira-Perpinan M., "Gaussian Mean-Shift Is an EM Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 767-776, 2007.

[3] Chang H. and Yeung D., "Robust Path-Based Spectral Clustering," *Pattern Recognition*, vol. 41, no. 1, pp. 191-203, 2008.

[4] Doan H. and Nguyen D., "A Method for Finding the Appropriate Number of Clusters," *The International Arab Journal of Information Technology*, vol. 15, no. 4, pp. 675-682, 2018.

[5] Fukunaga K. and Hostetler L., "The Estimation of The Gradient of A Density Function, with Applications in Pattern Recognition," *IEEE*

Transactions on Information Theory, vol. 21, no. 1, pp. 32-40, 1975.

[6] Hinneburg A. and Gabriel H., "DENCLUE 2.0: Fast Clustering Based on Kernel Density Estimation," *in Proceedings of International Symposium on Intelligent Data Analysis Berlin*, Ljubljana, pp. 70-80, 2007.

[7] Hinneburg A. and Keim D., "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," *in Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp. 58-65, 1998.

[8] Hubert L. and Arabie P., "Comparing Partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193-218, 1985.

[9] Kalti K. and Mahjoub M., "Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm," *The International Arab Journal of Information Technology*, vol. 11, no. 1, pp. 11-18, 2014.

[10] Khader M. and Al-Naymat G., "An Overview of Various Enhancements of DENCLUE Algorithm," *in Proceedings of the Second International Conference on Data Science, E-Learning and Information*, Dubai, pp. 1-7, 2019.

[11] Li X., Hu Z., and Wu F., "A Note on the Convergence of the Mean Shift," *Pattern Recognition*, vol. 40, no. 6, pp. 1756-1762, 2007.

[12] Luo Y., Zhang K., Chai Y., and Xiong Y., "Multi-Parameter-Setting Based on Data Original Distribution for DENCLUE Optimization," *IEEE Access*, vol. 6, pp. 16704-16711, 2018.

[13] Milligan G. and Cooper M., "A Study of The Comparability of External Criteria for Hierarchical Cluster Analysis," *Multivariate Behavioral Research*, vol. 21, no. 4, pp. 441-458, 1986.

[14] Muller E., Assent I., Gunnemann S., and Seidl T., "Scalable Density-based Subspace Clustering," *in Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, New York, pp. 1077-1086, 2011.

[15] Qiu R., Wang K., Li S., Dong J., and Xie M., "Big Data Technologies in Support of Real Time Capturing And Understanding of Electric Vehicle Customers Dynamics," *in Proceedings of the 5th International Conference on Software Engineering and Service Science*, Beijing, 2014.

[16] RenY., KamathU., Domeniconi C., and Zhang G., "Boosted Mean Shift Clustering," *in Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part II*, Nancy, pp. pp 646-661, 2014.

[17] Schneider J. and Vlachos M., "Fast Parameterless Density-based Clustering via Random Projections," *in Proceedings of the 22Nd ACM International Conference on Information and Knowledge Management*, New York, pp. 861-866, 2013.

[18] Yeung K. and Ruzzo W., "Details of the Adjusted Rand Index and Clustering Algorithms, Supplement to The Paper An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data, " *Bioinformatics*, vol. 17, no. 9, pp. 763-774, 2001.

[19] Zaki M., Meira W., and Meira W., *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014.

Mariam Khader She is currently working as a lecturer in Princess Sumaya University for Technology (PSUT), Amman, Jordan. She received the BSc degree in computer networking systems from the World Islamic Science & Education University (WISE) in 2012, Amman, Jordan. She received her MSc Degree in IT security and digital criminology in 2014 from PSUT. Currently, she is a PhD Candidate in computer science at PSUT. Between 2012-2015, she worked a teacher assistant and then a lecturer at the network department in the World Islamic Science and Education University. Her interests include digital forensics, network security and big data analytic.

Ghazi Al-Naymat He received his Ph.D. degree in May 2009 from the School of Information Technologies at The University of Sydney, Australia. He is currently working as an Associate Professor at the College of Engineering and Information Technology at Ajman University, UAE. In 2015, he joined the Department of Computer Science, King Hussein School of Computing Sciences at Princess Sumaya University for Technology (PSUT). He is a member of The Australian Computer Society. His research interests include Data Mining and machine learning, big data, and data science. Al-Naymat always targets reputable venues for his publications.