Dark Patches in Clustering

Waqar Ishaq, Eliya Buyukkaya
Computer Engineering Department
Kadir Has University
Istanbul, Turkey
waqar.ishaq@stu.khas.edu.tr, eliya.buyukkaya@khas.edu.tr

Abstract—**This survey highlights issues in clustering which hinder in achieving optimal solution or generates inconsistent outputs. We called such malignancies as dark patches. We focus on the issues relating to clustering rather than concepts and techniques of clustering. For better insight into the issues of clustering, we categorize dark patches into three classes and then compare various clustering methods to analyze distributed datasets with respect to classes of dark patches rather than conventional way of comparison by performance and accuracy criteria, because performance and accuracy may provide misleading conclusions due to lack of labeled data in unsupervised learning. To the best of our knowledge, this prime feature makes our survey paper unique from other clustering survey papers.**

Keywords— Clustering issues, taxonomy of clustering methods and model, clustering survey.

I. Introduction

The term *cluster* does not have a universal definition. Different researchers define cluster in various ways. This is due to different factors involved in the analysis (e.g. data type, application type, data size, domain expert, user), what expert and user anticipate from the data [2,5]. Besides, proximity problem occurring due to clusters of various shape and size with objects having different degree of belongingness may add further doubts to the evaluation step of the analytical model, making cluster definition ambiguous [5]. The main task in clustering is to discover sensible patterns in dataset. Clustering represents set of clusters, each contains homogeneous data objects. These objects may have degree of relationship with respect to hard clustering (i.e., an object either belongs to cluster or not) and soft clustering (i.e., an object belongs to two or more clusters based on certain degree of relationship) [8].

Nowadays, data are on move from legacy systems to distributed heterogeneous environment with multi-featured sources (e.g. IoT, sensors, etc.). Since data may have hidden relations in addition to its distributed characteristics, we use the term *federated* dataset to represent dataset within/among organizations. The possible hidden characteristics of federated data need to be considered to make proper clustering decision in heterogeneous environment.

In this survey, we focus on unsupervised learning rather than supervised learning because unsupervised learning can exploit freely relations among federated datasets due to non-reliance on the external information, whereas supervised learning confines its analytical model to class labels which may mask important features in federated environments [1,2,3].

Current clustering techniques don't produce considerable results for distributed heterogeneous environment, but have the potential to deliver promising outcomes by tinkering data analytics design procedure to draw inferences about distributed data through advance clustering algorithms. Before data are subjected to clustering, raw data are purified by passing through number of phases (e.g. pre-process used to deal with poor data) [9,10]. Such phases are beyond the scope of this survey. This survey focuses on the issues relating to clustering rather than concepts and techniques of clustering. For better understanding, clustering issues, called *dark patches*, are categorized into three, i.e., issues based on properties, evaluation and optimization criteria. To the best of our knowledge, the prime feature of this survey paper is to evaluate different clustering methods using classes of dark patches instead of performance and accuracy as evaluation criteria due to lack of labeled data.

The rest of the paper is organized as follows. Section 2 elaborates dark patches in clustering. Section 3 explains issues in clustering methods. Section 4 compares different clustering methods. Finally, section 5 concludes this work.

II. Dark Patches

Different clustering methods are used to analyze massive data. Clustering is not only ill-defined but also ill-posed problem, demanding some prior knowledge of decent analysis [2]. To identify true patterns and their behaviors with high performance and accuracies within the constraints of application/organization is challenging task in heterogeneous environment. The reasons behind clustering issues are methodology limitations, evaluation misinterpretation, lack of knowledge about applications and possible hidden relations among federated data, which cause performance degradation, result inaccuracies and ambiguity in data analysis. There is no single algorithm dealing with all types of clustering structures and properties [11]. Moreover, resulting output of clustering may be artifact of the clustering method itself rather than representing true data structure. We use the term *dark patches* to represent such issues in clustering which lead to hinder in achieving optimal solution or generate inconsistent outputs. We divide dark patches of clustering into three categories: issues based on clustering properties, evaluation and optimization

978-1-5386-0930-9/17/$31.00 ©2017 IEEE

criteria which directly, indirectly or both may malign the final clustering outputs.

A. *Issues based on Properties*

Data clustering groups objects such that there is dissimilarity among groups (clusters) whilst similarities among objects within same group for the sake of pattern discovery which helps to project data from high dimensional to low dimensional space. However, issues related to clustering properties such as *cluster initialization, number of clusters, empty cluster, outliers, local minima, grey zones, non-spherical clusters and clusters of various sizes* are shadows in clustering methods to malign data analytics [12].

The violation of clustering properties by clustering model under certain conditions may cause erroneous outputs like inconsistency, inaccuracy with low quality [4]. For unsupervised learning, it is not possible to have prior knowledge about true number of clusters. If number of selected clusters is less than true number of clusters for a dataset, then diverse objects will lie within same cluster, causing object masking. To deal with such problem, minimum square error criteria is used for convex shape clusters, which does not work for arbitrary shape clusters [13]. Different clustering methods may produce incompatible data results when number of clusters are different [4]. Empty cluster is another issue increasing squared error which requires centroid replacement strategy to maintain square error, but works under certain limits [12]. In addition, distance which has many types such as distance between instances, distance between instance and cluster centroids, distance between centroids of clusters, is used in many clustering algorithms to organize data into groups. However, different distance types may lead to various proximity problems [2]. Moreover, to compare the results of different clustering methods for same dataset is not straightforward due to different number of clusters, variation in cluster sizes and their boundaries. Problem becomes even worse if dataset has high dimensionality [30]. Furthermore, centroid initialization issue involves excessive iterations if selected centroid is not close to true cluster centroid. Overlapping area among two or more clusters results in grey zone. The larger the overlapping area, the more the results will be prone to erroneous outputs [14].

B. *Issues based on Evaluation Criteria*

Since clustering algorithms are based on different assumptions about cluster shape, similarity matrix and grouping criteria for unlabeled data, there do not exist universal cluster evaluation criteria to evaluate all clustering algorithms, which makes clustering validation more challenging [12,15]. Indexes are designed to measure validity of clustering results. However, they may favor one solution over another due to arbitrary cluster shape. Therefore, domain expert must evaluate the output and modify the assumptions to look for the optimal solution [2]. Validity indices are mostly suitable for crisp clustering where there is no overlapping [12,15]. To describe the goodness of clustering, cluster validity uses compactness and separation as parameters to measure quality of clustering, also called similarity matrix. These two parameters have inverse co-relation, i.e. when one improves other deteriorates. Therefore, relative weights are used to balance the model which makes the selected model biased [2]. The main disadvantage of validity indices is that they cannot measure arbitrary shape cluster. Affinity parameter is used for spectral cluster which is beyond the scope of this survey. There are many cluster validity schemes [15,16], e.g. Dunn and Dunn like indices, Davies Bouldin Index, Root Mean Square Standard Deviation, Root Squared, SD validity index and S_Dbw Validity Index. They have different approaches in clustering evaluation but have common issues such that time complexity is O(n^2), they are not valid for clusters of arbitrary shapes, right clustering scheme cannot be identified without well separated clusters due to highly dense data, etc. Besides, Dunn and Dunn is also sensitive to outlier. Silhouette Index is also internal criteria for clustering analysis which works well with highly dense data but does not work with arbitrary shapes and has O(n^2). There are additional cluster evaluation methods like quantization error [16] and diversity [17] approaches in case of neural networks.

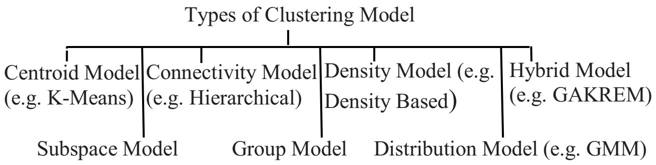
Types of Clustering Model
Centroid Model (e.g. K-Means)
Connectivity Model (e.g. Hierarchical)
Density Model (e.g. Density Based)
Hybrid Model (e.g. GAKREM)
Subspace Model
Group Model
Distribution Model (e.g. GMM)

Fig. 1. Taxonomy of clustering models

C. *Issues based on Optimization Criteria*

With the exponential growth of data, traditional approaches are not capable to counter technological challenges, causes shift in paradigm from traditionally centralized data with centralized processing, to distributed data with distributed processing. Like scalability issue, computational constraint, resources limitations and response time constraints are few among many reasons for such data migration [18,19]. Different clustering algorithms are used to extract useful information to make intelligent decision integration in federated data environment. In centralized system, data pooling is infeasible to perform clustering due to distributed nature of data across different databases resulting in high computational cost [18].

Massive data with high dimensionality having wide and tall datasets, have issues of multi-dimensional feature visualization and data resolution, respectively. To process tall data, we need scalable distributed systems which may result in stale operations to execute long queries. Likewise, to analyze wide datasets, multiplexing of various features may further complicate the designed model [20,21]. This process of transforming and combining diverse data is difficult and time consuming task. Besides tall and wide datasets issues, to manipulate large volumes of data interactively within time constraint is another issue which may result in drift concept [21].

Since there is no universal clustering algorithm that satisfies all needs of user [18], we need integrated approach to tackle problems of large complex heterogeneous data residing in federated environment. Therefore, clustering architectures and models of various types are designed to manage giant data [8,12,23,24]. Fig. 1 shows the taxonomy to elaborate clustering models.

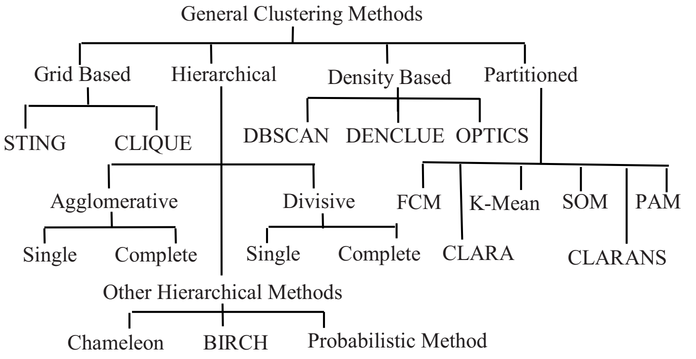

Fig. 2. Taxonomy of general clustering methods

Global clustering (GC), local clustering without collaboration (LC) and local clustering with collaboration (LCC) are three approaches used to find the analytical differences among different cluster architectures [19]. According to GC, entire data are pooled for analysis. In case of LC, patterns are developed in distributed environment without sharing data and results. Whereas LCC has no data pooling but shares results among federated datasets.

Consequently, different clustering algorithms are developed to deliver proper solutions in diverse environment with optimal performance. But with the increase in the number of methods to analyze data, the size of search space also increases, leading to many local optimal. Additionally, number of features and their (ir)relevance have strong impact on final solution, because with increase in dimension, search space increase exponentially causing data sparsity [2,18]. Clustering optimization criteria involves many local optimal solutions which requires many parameters to be tuned and adjusted by domain expert [2].

III. Clustering Methods and Their Issues

Clustering methods are used to develop groups of objects for the given dataset based on their clustering properties. There is no universal clustering method to discover hidden patterns for multi-dimensional data as an optimal solution, but there are certain requirements for better clustering. Therefore, various methods of clustering are used to meet these requirements to get optimal solution. We categorize clustering methods into two sections, general clustering and combinational clustering methods, to elaborate issues corresponding to each method.

A. General Clustering Methods

There are many clustering methods to examine data. Since highlighting all clustering methods is not possible, we draw Fig. 2 to describe the taxonomy of general clustering methods by reviewing [3]. General clustering methods provide single solution for given datasets. Each method has its own pros and cons based on its assumptions. As data are unlabeled, different clustering methods may produce different results for same data. This may be due to limitations of each clustering method and its properties. Table I-IV display issues related to each clustering method, which are then used to compare different methods based on categories of dark patches in Table V.

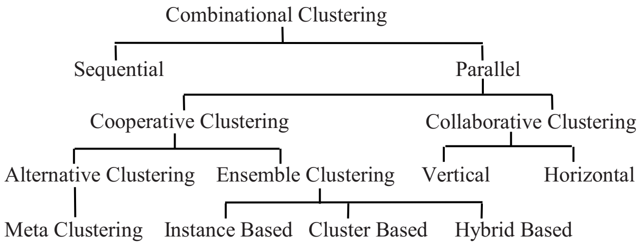

Fig. 3. Taxonomy of combinational clustering

B. Combinational Methods of Clustering

Traditional clustering methods emphasis on single solution even though there may exist multiple solutions. These multiple solutions are due to either ill-define nature of clustering or high dimensional data. Likewise, clustering may be good for one criterion, but can be suboptimal for another criterion [16,18]. This creates a biased solution because of either clustering method's artifact or data masking. Moreover, it is not possible to meet all clustering properties by single clustering method [18]. Therefore, we need to combine different clustering algorithms to get general purpose optimal solution to discover dissimilar alternate clustering. Combinational clustering, also called hybrid clustering, has certain complexities like clustering overhead, jeopardizing similar clustering, inefficiency of running algorithms multiple times and (dis)similar clustering solution evaluation [16]. We conclude by constructing Fig. 3 to show taxonomy of combinational clustering methods [16,18].

1) Sequential clustering: According to this approach, there is cascaded arrangement of clustering algorithms to generate optimal output such that output of previous method is input to next clustering method. [31] describes fast DBSCAN method where K-mean and DBSCAN are used collectively to analyze large dataset, and provides promising results in match to general DBSCAN. Fast DBSCAN with its variants provides competitive results but it is sensitive to data sequence and has synchronization issue. If acquired data are real time, then clustering must be done in real time to produce consistent results, which is challenging. According to [33], clustering seeking method is used to associate samples with the nearest cluster unless distance is greater than certain threshold to generate new cluster. This method is sensitive to input order and threshold level. These issues may be fixed by setting multiple thresholds, called certainty in cluster belongingness. But such variants give rise to convergence problem. [34] highlights feature based vs similarity based sequential clustering approaches to project sequential data into frequent patterns using K-means clustering. Feature based approach gives better solution but it is difficult to transform sequential data into sequential feature space, because there may be high dependencies among features corresponding to particular data sequence to perturb the similarity measurements. [35] combines genetic algorithm (GA) and logarithmic regression with K-means and EM algorithm, called GAKREM (genetic algorithm K-means logarithmic regression expectation maximization) to overcome drawbacks of parameter initialization and determining number of clusters. GAKREM combines best characteristics of K-means and EM algorithms, avoids their weaknesses, but has optimization issue due to synchronization.

2) Parallel clustering: According to this approach, data are executed simultaneously by various clustering methods to get optimal solution. Here, clustering methods are adjusted to overcome the deficiencies of each method. Parallelization clustering methods have issues of optimal solution and synchronization [11]. [32] explains parallelization approach by using K-means method in Apache Storm and but has synchronization and high dimensionality sparsity issues. Moreover, variation in arrival rates and latency in distributed environment for real time application may create drift concept. Following is the classification of parallelization clustering to elaborate their issues in detail:

a) Cooperative clustering: The concept behind cooperative clustering is to achieve common benefits by establishing consensus among different algorithms. This consensus is used to combine outputs of different individual clustering algorithms via some control algorithm called master algorithm. For unbiased output, master algorithm must confirm symmetry and impartial combination of individual algorithms by maintaining properties of each individual algorithm to larger extent.

i)Alternative clustering: The idea behind alternative clustering is that there exists more than one clustering solution for given datasets. This is due to either high dimensional data or complex data that cannot be explained by single solution. This requires object of multi-nature to provide multiple solutions rather than unimodal objective function [18]. Dual objective function, meta clustering, Markov Chain Monte Carlo (MCMC) approach and data transformation into subspace are four approaches used in alternative clustering. But these approaches are methodology specific, are not suitable for large datasets, have time constraint and are not applicable to low feature space respectively [18]. Constrained Orthogonal Average Link Algorithm (COALA) is alternative clustering using hierarchical clustering algorithm with can/cannot link constraints as pre-requisite to find new clustering with high quality and dissimilarity from other clustering [16]. Here, the drawback is that quality degrades by imposing cannot link constraint as pre-requisite which depletes important relations within dataset. Likewise, it requires multiple runs of clustering algorithms to discover patterns [26]. Multiple run clustering system (MRCS) is alternative clustering approach that selects arbitrary clustering algorithm parameters automatically, but size of search space increases with increase in number of different algorithms which cause optimization issue [18]. Meta Clustering is alternative clustering where there is interaction among users, cluster system and data to find many alternate solutions for given data and then allows user to navigate for the best clustering [18]. But this approach has clustering initialization and optimal solution navigation problems [18].

ii) Ensemble clustering: Ensemble clustering, also called consensus or aggregation clustering, is a powerful tool to handle traditional issues of clustering by combining multiple algorithms. The aim is to generate multiple clustering and then merge clusters to find final consensus clustering solution. By such approach, one or more aspects of clustering algorithms, parameter values and number of features or objects are adjusted to generate unbiased outputs [4,6,27]. Ensemble clustering is effective for distributed computing environment if results are generated independently [6]. Instance-, cluster- and hybrid-based ensemble clustering are three categories of ensemble clustering having certain issues in common. Consensus function considers all clustering ensemble solutions equally without considering correlation among different clusters. This causes biased consensus function. Moreover, there are issues of diversity measurement and weighted consensus clustering problem [4]. Such biased ingredients are handled by weight based approaches applied to each of the above categories of ensemble clustering, e.g. weighted instance based clustering ensemble (WICE), etc. But such clustering ensemble solutions are still unfair to produce desire results due to issue of diversity where there exists correlation among clustering ensemble solutions [3].

b) Collaborative clustering: Collaborative clustering processes datasets locally and then collaborates with remote sites about their discoveries, involving two steps, i.e. local phase and collaboration phase [17,19,28]. There are two main categories of collaborative clustering to discover data patterns i.e. vertical and horizontal collaborative clustering [2,17,28]. Each type has its own issues and concerns. But horizontal clustering is more challenging due to group of objects belonging to different feature space. Those major issues which create hindrance in the full adaptation of collaborative approach for knowledge discovery are selection of collaborators, conflict management, learning mechanism among data sites, noise and global feature assessment [16,29]. Since there is no direct link among clustering results during collaboration phase at each data site, conflicts may arise among results due to either different types of algorithms used to execute same data or change in parameter using similar algorithms for same datasets [17,22]. Datasets are not noise free. Noisy features in datasets are random and less informative but are used to indicate relevant information in datasets. During collaboration, lack of related features is not only due to noise features but also unrelated objects in the feature space, are sources of noise. Thus, global feature assessment in the federated environment is complex due to presence of noise and unrelated features. So, collaborative matrix and confidence links are used to establish relations among federated datasets to cope with above challenges [17,22,28,29].

IV. Comparison of Different Clustering Methods

Since there is no universal approach to measure performance and accuracy of unsupervised datasets due to ill-defined nature of clustering and lack of class labeling, we compare different clustering methods based on classes of dark patches. This provide us common platform for comparison rather than using conventional performance and accuracy parameters.

V. Conclusion

General clustering methods do not respond well in terms of performance and accuracy in distributed heterogeneous environment with multi-featured sources. This may be due to methodology limitations, evaluation misinterpretation, lack of knowledge about data applications and hidden relations among federated data. Combinational clustering methods have caliber to overcome the limitations of general clustering methods to improve quality of analytics by executing data simultaneously in distributed environment. But this requires common platform to manipulate and validate dependencies among federated datasets to produce global structure. Clustering dark patches highlight reasons of clustering inaccuracies and performance evaluation degradation to extract true data patterns in federated environment.

TABLE I. Partitional Clustering Methods

Method	Drawbacks	Remarks
K-Mean	1. Global minimum not guaranteed 2. Number of clusters to be known 3. Requires multiple runs for better results 4. Sensitive to outlier 5. Complexity O(nkt)	Variants exists
PAM	1. Computational cost is high $O(k(n-k)^2)$ 2. Number of clusters to be known 3. Unfit for large dataset	Reduced MSE
CLARA	1. Computational cost is $O(ks^2+k(n-k))$ 2. No best clustering for bad sampling 3. Depends upon size of samples	Efficiency enhanced by sampling
CLARANS	1. Computational cost increase by L 2. Depends upon sample size	Randomizing samples L times
FCM [8,12,17,24]	1. Not right choice for densely object 2. Do not guarantee optimal cluster centers	Variant of K-mean Better choice for noisy data^{44}
SOM [7,30]	1. Require large time to train system 2. No specific way to initialize parameters	Perform clustering and visualization simultaneously

TABLE II. Hierarchical Clustering Methods

Method	Drawbacks	Remarks
Agglomerative Or Divisive Single/complete link	1. Needs prior knowledge about number of clusters 2. Both are over sensitive to noise ~ use mean 3. Divisive method is not good for large datasets 4. In divisive, backtrack on partition not possible 5. Issue of scalability 6. Time complexity is O (n^2log n)	More challenges in divisive, but agglomerative has more variants
Other methods (BIRCH)	1. Computational cost is O(n) 2. Don't perform well for non-spherical clusters. 3. It is order sensitive [8]	Adds scalability by integrating with other clustering methods
Other methods (Chameleon)	1. Computational cost is $O(n^2)$	Ability to deal with arbitrary shape clusters
Other methods (Probabilistic)	1. Can't handle uncertainty of clustering hierarchies	Overcome drawbacks of hierarchical method

TABLE III. Density Based Clustering Methods

Method	Drawbacks	Remarks
DBSCAN	1. Setting parameters manually is problematic for high dimensional data 2. Time complexity for spatial index is O(nlogn) else $O(n^2)$ 3. Monotonic approach (fixed values)	To find clusters of arbitrary shapes
OPTICS	1. Time complexity for spatial index is O(nlogn) else $O(n^2)$ 2. Sensitive to radius for density measurement 3. Variation is density	Do not require density threshold
DENCLUE	1. Time complexity is O (nlogn) 2. Hill climb ineffective to get optimal solution	Use non-parametric method to deal with density estimation

TABLE IV. Grid Based Clustering Methods

Method	Drawbacks	Remarks
STING	1. Time complexity is O(n) and query processing time is O(g), g is grid cell 2. Accuracy degrades due to isothetic nature of cluster	Enhance processing speed
CLIQUE	1. Accuracy falls due to grid size and density threshold 2. Dense projections with high features have dense overlaps	Find density based clusters in subspace with non-overlapping partitions

TABLE V. General Clustering Methods Comparison

Categories	Method	Architecture	Model	Dark Patches Based On		
				Properties	Evaluation	Optimization
Partitional Methods	K-Mean [25]	Global Clustering or Local Clustering	Centroid	✓	✓	✓
	PAM			✓		✓
	CLARA			✓	✓	
	CLARANS				✓	
	FCM [8,12,19,24]				✓	✓
	SOM [7,30]			✓		✓
Hierarchical Methods	Agglomerative Or Divisive with Single/complete link		Connectivity		✓	✓
	Other methods (BIRCH)				✓	
	Other methods (Chameleon)					✓
	Other methods (Probabilistic)					✓
Density Based	DBSCAN		Density	✓	✓	
	OPTICS				✓	✓
	DENCLUE					✓
Grid Based	STING				✓	
	CLIQUE				✓	

TABLE VI. COMBINATIONAL CLUSTERING METHODS COMPARISON

Categories	Method	Architecture	Model	Dark Patches Based On		
				Properties	Evaluation	Optimization
Sequential Clustering	fast DBSCAN	Global Clustering	Hybrid		✓	✓
	Clustering seeking				✓	✓
	GAKREM					✓
Parallelization	Hierarchical Parallelization					✓
	K-Means Parallelization				✓	✓
	Alternative clustering	Local Clustering without Collaboration		✓		✓
	Meta clustering			✓	✓	✓
	Ensemble clustering			✓	✓	
	Vertical Collaborative clustering	Local Collaborative Clustering		✓	✓	✓
	Horizontal Collaborative clustering			✓	✓	✓

REFERENCES

[1] A.K. Jain, M.N. Murty and P.J.Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, no. 3, pp. 262-323, Sept. 1999.

[2] A. Cornueejols, C. Wemmert, P. Gancarski and Y. Bennani, "Collaborative Clustering: Why, When, What and How", International Journal on Information Fusion, Elsevier, vol. 39, pp. 81-95, Jan. 2018.

[3] J. Han, M. Kamber and J. Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, Elsevier, 2012.

[4] A. Topchy, A. K. Jain and W. Punch, "Clustering ensembles: models of consensus and weak partitions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

[5] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", pp. 320, Prentice-Hall, Inc. ISBN 0-13-022278-X, 1988.

[6] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.

[7] M. Ghassany, N. Grozavu and Y. Bennani, "Collaborative Multi-View Clustering", in Proc. IJCNN, IEEE International Joint Conference on Neural Network, Dallas, TX- August 4-9, 2013.

[8] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques", Journal of Intelligent Information Systems, vol. 17, num. 2, pp. 107-145, Dec. 2001.

[9] F. Rashid, A. Miri and I. Wougang, "A Secure Video Deduplication Scheme in Cloud Storage Environments using H.264 Compression," Proc. of the 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015.

[10] M. Balazinska, A. Deshpande, M. J. Franklin, P. B. Gibbons, J. Gray, S. Nath, M. Hansen, M. Liebhold, A. Szalay and V. Tao, "Data Management in the Worldwide Sensor Web," in IEEE Pervasive Computing, vol. 6, no. 2, pp. 30-40, April-June 2007.

[11] K., Rasha, and M. S. Kamel, "Cooperative clustering." Pattern Recognition vol. 43, num. 6, pp. 2315-2329, 2010.

[12] P.-N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining", Pearson, 2006.

[13] A. Fred, "Finding Consistent Clusters in Data Partitions", Proc. Of Multiple Classifier Systems: Second International Workshop, pp. 309-318, July 2001.

[14] Y. Zeng, J. Tang, J. Garcia-Frias and G. R. Gao, "An Adaptive Meta-Clustering Approach: Combining the Information from Different Clustering Results", Proc. of the IEEE Computer Society Conference on Bioinformatics, pp. 276, Aug 2002.

[15] F. Kovacs, C. Legany and A. Babos, "Cluster Validity Measurement Techniques", Proc. of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, pp 388-393, Feb 2006.

[16] E. Bae and J. Bailey, "COALA: A Novel Approach for the Extraction of an Alternate Clustering of High Quality and High Dissimilarity", Sixth International Conference on Data Mining, pp.53-62, 2006.

[17] P. Rastin, G. Cabanes, N. Grozavu and Y. Bennani, "Collaborative Clustering: How to Select the Optimal Collaborators?" 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, pp. 787-794, 2015.

[18] R. Caruana, M. Elhaway, N. Nguyen and C. Smith, "Meta Clustering", Proc. of the Sixth International Conference on Data Mining, pp. 107-118, Dec. 2006.

[19] B. Depaire, R. Falcon, K. Vanhoof, and G. Wets, "Pso driven collaborative clustering: A clustering algorithm for ubiquitous environments," Intell. Data Anal., vol. 15, no. 1, pp. 49–68, Jan. 2011.

[20] K. Grolinger, M. Hayes, W. A. Higashino, A. L. Heureux, D. S. Allison and M. A. A. Capretz, "Challenges for MapReduce in Big Data"

[21] J. Heer and S. Kandel, "Interactive Analysis System"

[22] G. Forestier, C. Wemmert and P. Gancarski, "Towards conflict resolution in collaborative clustering," 5th IEEE International Conference Intelligent Systems, pp. 361-366, London, 2010.

[23] K. Hammouda, M. Kamel, "Collaborative Document Clustering" Proc. of the SIAM International Conference on Data Mining, 2006.

[24] .A. Ben Ayed, M. Ben Halima and A. M. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data," 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, pp. 331-336, 2014.

[25] S. Bettoumi, C. Jlassi and N. Arous, "Comparative Study of K-means Variant for Mono-view Clustering", 2nd International Conference on Advanced Technologies for Signal and Image Processing, Monastir, Tunisia, March 2016.

[26] R. Jiamthapthaksin, C. F. Eick, and V. Rinsurongkawong, "An Architecture and Algorithms for Multi-Run Clustering" Computational Intelligence and Data Mining, IEEE Symposium, 2009.

[27] F. Gullo, A. Tagarelli and S. Greco" Diversity-based Weighting Schemes for Clustering Ensembles", Proc. of the 2009 SIAM International Conference on Data Mining, pp. 437-448, 2009.

[28] M. Ghassany, N. Grozavu, and Y. Bennani, "Collaborative clustering using prototype-based techniques," International Journal of Computational Intelligence and Applications, vol. 11, no. 03, p. 1250017, 2012.

[29] N. Grozavu, M. Ghassany, and Y. Bennani, "Learning confidence exchange in collaborative clustering," in Neural Networks (IJCNN), The International Joint Conference on, pp. 872–879, 2011.

[30] C. K. Fong, "A Study in Deploying Self-Organized Map (SOM) in an Open Source J2EE Cluster and Caching System," IEEE/ICME International Conference on Complex Medical Engineering, Beijing, pp. 778-781, 2007.

[31] V. V. Thang, D.V. Pantiukhin and A.I. Galushkin, "A Hybrid Clustering Algorithm: the Fast DBSCAN", V. V. Thang, D. V. Pantiukhin and A. I. Galushkin, "A Hybrid Clustering Algorithm: The FastDBSCAN," Int. Conf. on Engineering and Telecommunication, pp. 69-74, 2015.

[32] X. Gao, E. Ferrara and J. Qiu, "Parallel Clustering Of High-Dimensional Social Media Data Streams" Cluster, Cloud and Grid Computing, 15th IEEE/ACM International Symposium on, pp. 323-332, 2015.

[33] P. Trahanias and E. Skordalakis, "An efficient Sequential Clustering Method", Journal Pattern Recognition, vol. 22, num. 4, pp. 449-453, 1989.

[34] V. Guralnik and G. Karypis, "A Scalable Algorithm for Clustering Sequential Data", 1st IEEE Conference on Data Mining, pp.179-186, 2001.

[35] C. D. Nguyen and K. J. Cios, "GAKREM: A Novel Hybrid Clustering Algorithm", Inf. Sci., vol. 178, pp. 4205-4227, 2008.