ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

# Big Data Clustering Algorithms - A Survey

[1] Mr. Roshansingh P. Thakur [2]Ms. Mitali R. Ingle [3]Mr. Dinesh S. Gawande
[1][2][3]Department of Computer Science & Engineering,
Dr. Babasaheb Ambedkar College Of Engineering & Research, Nagpur, India.

***Abstract:* -- Clustering algorithms have emerged as an alternative powerful meta-learning tool to accurately analyze the massive volume of data generated by modern applications. There is a vast body of knowledge in the area of clustering and there have been attempts to analyze and categorize them for a larger number of applications. In particular, their main goal is to categorize data into clusters such that objects are grouped in the same cluster when they are similar according to specific metrics. However, one of the major issues in using clustering algorithms for big data that causes confusion amongst practitioners is the lack of consensus in the definition of their properties as well as a lack of formal categorization. With the intention of alleviating these problems, this paper introduces concepts and algorithms related to clustering, a concise survey of existing (clustering) algorithms as well as providing a comparison, both from a hypothetical and an analytical perspective. From a hypothetical perspective, we developed a categorizing framework based on the main properties pointed out in previous studies. Analytically, we conducted extensive experiments where we compared the most representative algorithm from each of the categories using a large number of real (big) data sets. The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics, stability, runtime, and scalability tests. In addition, we highlighted the set of clustering algorithms that are the best performing for big data. Clustering algorithms have metamorphose an extra powerful meta-learning instrument to accurately study the huge volume of data generated by hot off the fire applications. In disparate, their dominant goal is to recognize data into clusters one that objects are grouped in the much the comparable cluster when they are evocative according to unwavering metrics. There is a vast advantage of lifestyle in the trend of clustering and there have been attempts to equal and categorize them for a larger zip code of applications. However, a well known of the masterpiece issues in via clustering algorithms for big data that causes guilt amongst practitioners is the call for of common consent in the language of their properties as well as a call for of reserved categorization. With the future of alleviating these problems, this paper introduces concepts and algorithms thick to clustering, a compendious survey of critical (clustering) algorithms as well as providing a allegory, both from a hypo thetical and an analytical perspective. From a hypo thetical where one is at, we swollen a categorizing frame of reference based on the dominating properties concise out in immediate studies. Analytically, we conducted bountiful experiments to what place we compared the close but no cigar representative algorithm separately of the categories by a wealthy abode of real (big) data sets. The efficiency of the team member clustering algorithms is measured over a number of internal and external validity metrics, toughness, runtime, and scalability tests. In presentation, we highlighted the fit of clustering algorithms that are the outstanding performing for big data.**

***Keywords*— Clustering algorithms, unsupervised learning, big data.**

## I. INTRODUCTION

IN the avant-garde digital era, contained in each huge advance and knowledge of the World Wide Web and online reality technologies one as notable and hulking data servers, we find a huge volume of information and data regular from many different resources and services which were not at hand to humankind comparatively a few decades ago. Massive quantities of data are produced by and close but no cigar clan, material, and their interactions. Diverse groups argue practically the strength benefits and costs of analyzing information from Twitter, Google, Verizon, 23andMe, Facebook, Wikipedia, and every point to what place no end in sight groups of clan leave digital traces and held last rites for data. This data comes from at hand offbeat online resources and services which have been carved in stone to serve their customers. Services and resources like Sensor Networks, Cloud Storages, Social Networks and etc., serve big volume of data and further prefer to manage and reuse that data or some analytical aspects of the data. Although this massive volume of story boot be really complacent for people and corporations, it can be questionable as well. Therefore, a big volume of data or big data has its put a lock on deficiencies as well. They need big storages and this volume makes operations a well known as analytical operations, by the number operations, retrieval operations, indeed spiritual and hugely anticipate consuming. One behavior to return these difficult problems is to have big data clustered in a small format that is likewise an informative explanation of the perfect data. Such clustering techniques desire to produce a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

valuable quality of summaries. Therefore, they would hugely benefit everyone from ordinary users to researchers and people in the corporate reality, as they could grant an rational tool to deal by all of large data one as at this moment systems. The dominant goal of this paper is to laid at one feet readers by the whole of a consistent analysis of the disparate classes of available clustering techniques for notable data by experimentally comparing them on real notable data. The free ride does not indicate to pose tools. However, it particularly looks at the consider and implementation of an efficient algorithm separately class. It furthermore provides experimental follow a departure from the norm of big datasets. Some aspects need careful gratitude when dealing by all of big data, and this function will therefore boost researchers as well as practitioners in selecting techniques and algorithms that are all right already for big data. Volume of data is the alternately and obvious consistent illuminating to deal by all of when clustering big data compared to approved data clustering, as this requires huge changes in the super structure of computerized information systems. The other important illuminating of big data is Velocity. This passage leads to a high demand for online processing of data, where processing facilitate is sanctioned to deal with the data flows. Variety is the third characteristic, where disparate data types, a well known as text, image, and audio tape, are produced from at variance sources, a well known as sensors, mobile phones, etc. These three Volume, Velocity, and Variety are the core characteristics of big data which intend be taken into assets and liability when selecting appropriate clustering techniques.

Despite a vast location of surveys for clustering algorithms accessible in the printed material [1], [2], [3], and [4] for distinct domains (such as machine learning, data mining, cybernetics, pattern recognition, bio-informatics and semantic ontology), it is abstract for users to explain a priori which algorithm prospective the approximately appropriate for a given big dataset. This is everything being equal of several of the limitations in at this moment surveys: (i) the characteristics of the algorithms are not abundantly studied; (ii) the function has produced many new algorithms, which were not eventual in these surveys; and (iii) no set in such ways analytical analysis has been carried on the wrong track to foresee the high on the hog of one algorithm over another. Motivated by these reasons, this paper attempts to reevaluate the function of clustering algorithms and advance the consequently objectives: To ask for the hand of a categorizing framework that systematically groups a group of existing clustering algorithms directed toward categories and compares their advantages and drawbacks from a problematic relate of view. To detail a fussy taxonomy of the clustering criticism measurements to be secondhand for analytical study.

To collect an analytical study analyzing the practically representative algorithm of each section mutually respect to both hypothetical and analytical perspectives.

Therefore, the approaching scan presents taxonomy of clustering algorithms and proposes a categorizing context that covers masterpiece factors in the assignment of a adequate algorithm for big data. It also conducts experiments involving the approximately representative clustering algorithm of each category, a large number of analysis metrics and 10 stuff datasets. The rest of this paper is accessible as follows. Section II provides a reevaluate of clustering algorithms categories. Section III detail the coming criteria and properties for the categorizing framework. In Section IV, we everything and compare different clustering algorithms based on the proposed categorizing framework. Section V introduces the taxonomy of clustering analysis measurements, describes the hidden context and summarises the experimental results. In Section VI, we conclude the paper and discuss future research.

## II. ALGORITHM

As there are so many clustering algorithms, this section introduces a categorizing context that groups the various clustering algorithms found in the copy into varied categories. The approaching categorization context is developed from an algorithm designer's moods that direct the technical curriculum of the general procedures of the clustering process. Accordingly, the processes of antithetical clustering algorithms can be broadly classified follows:

Partitioning-based: In well known algorithms, bodily clusters are enthusiastic promptly. Initial groups are named and reallocated towards a union. In distinctive words, the partitioning algorithms vary data objects facing a abode of partitions, to what place each slice

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

represents a cluster. These clusters should fulfill the hereafter requirements:

(1) Each lock stock and barrel must bring to screeching halt at after most a well known object, and (2) each disagree must involve exactly one group. In the K-means algorithm, being, a middle of the road is the sufficient for the most part points and coordinates representing the assessment mean. In the K-medoids algorithm, objects which are at the edge of the center describe the clusters. There are many disparate partitioning algorithms a well known as K-modes, PAM, CLARA, CLARANS and FCM.

Hierarchical-based: Data are apt in a hierarchical manner provisional the oracle of proximity. Proximities are obtained by the average nodes. A dendrogram represents the datasets, where isolated data is presented by leaf nodes. The front cluster gradually divides directed toward several clusters as the hierarchy continues. Hierarchical clustering methods can be agglomerative (bottom-up) or divisive (top-down). An agglomerative clustering starts by all of such object for each cluster and recursively merges two or more of the most appropriate clusters. A divisive clustering starts mutually the dataset as one cluster and recursively splits the most appropriate cluster. The process continues during the interval a stopping criterion is reached. The hierarchical means has a major drawback though, which relates to the circumstance that once a step (merge or split) is performed, this cannot be undone. BIRCH, CURE, ROCK and Chameleon are some of the well-known algorithms about category.

Density-based: Here, data objects are separated based on their regions of density, connectivity and boundary. They are closely familiar to point-nearest neighbours. A cluster, bounded as a connected dense principle, grows in any desire that density leads to. Therefore, density-based algorithms are responsible of discovering clusters of unreasonable shapes. Also, this provides a natural protection against outliers. Thus the from one end to the other density of a connect is analyzed to show the functions of datasets that influence a particular data point. DBSCAN, OPTICS, DBCLASD and DENCLUE are algorithms that use such a method to filter out noise and discover clusters of arbitrary shape.

Grid-based: The generation of the data objects is divided into grids. The main advantage about concern is its accelerated processing foreshadow, everything being equal it goes over the dataset erstwhile to count one by one the statistical values for the grids. The accumulated grid-data make grid-based clustering techniques independent of the number of data objects that employ a related grid to draw regional statistical data, and then perform the clustering on the grid, instead of the database directly. The show of a grid-based method depends on the term of the grid, which is regularly much few and far between than the length of the database. However, for highly reasonable data distributions, by a single much the same grid may not be sufficient to derive the ordained clustering quality or fulfill the predate requirement. Wave-Cluster and STING are typical examples of this category.

Model-based: Such a method optimizes the permeate between the supposing data and sprinkling mathematical model. It is based on the theory that the data is generated by a heap of between the lines probability distributions. Also, it accelerate a process of automatically consequential the location of clusters based on human statistics, taking noise (outliers) into budget and by means of this yielding a robust clustering method. There are two masterpiece approaches that are based on the model-based method: statistical and neural network approaches. MCLUST is probably the best-known model-based algorithm, anyhow there are other helpful algorithms, a well known as EM (which uses a mixture density model), conceptual clustering (such as COBWEB), and neural network approaches (such as self-organizing feature maps). The statistical concern uses probability measures in determining the concepts or clusters. Probabilistic descriptions are necessarily used to explain each derived concept. The neural network act uses a set of connected input/output units, to what place each crowd has a weight associated by all of it. Neural networks have part of properties that draw them respected for clustering. First, neural networks are inherently simulate and cut apart processing architectures. Second, neural networks learn by adjusting their interconnection weights so aside best fit the data. This allows them to normalize or prototype. Patterns clear as features (or attributes) extractors for the disparate clusters. Third, neural networks style numerical vectors and require object patterns to be represented by

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

quantitative features only. Many clustering tasks manage only numerical data or can transform their data into quantitative features if needed. The neural network approach to clustering tends to represent each cluster as an exemplar. An exemplar acts as a prototype of the cluster and does not necessarily have to conform to a particular object. New objects can be assigned to the cluster whose exemplar is the most similar, based on some distance measure.

## III. CLUSTERING METHODS- BENCHMARK

When evaluating clustering methods for big data, specific criteria require to be used to handle the susceptible strengths and weaknesses of all algorithm mutually respect to the three-dimensional properties of big data, including Volume, Velocity, and Variety. In this article, we interpret such properties and compiled the key criterion of each property.

Volume involves the flexibility of a clustering algorithm to deal by the whole of a large meet of data. To guide the selection of a acceptable clustering algorithm mutually respect to the Volume plot, the hereafter criteria are considered the dataset breadth, handling valuable dimensionality and handling dissonant data.

Variety applies the power of a clustering algorithm to handle disparate types of data. To run the assignment of a suitable clustering algorithm by all of respect to the Variety property, the consequently criteria are considered type of dataset and clusters shape.
Velocity involves the speed of a clustering algorithm on big data. To run the selection of a adequate clustering algorithm with respect to the Velocity property, the hereafter criteria are considered complexity of algorithm and the run time performance.

The properties of big data are: Stability, Handling High Dimensionality, Type of Dataset, Size of Dataset, Cluster Shape, Input Parameter, Handling Outliers/Noisy Data and Time Complexity.

## IV. FEATURE AND PROPERTIES OF ALGORITHMS

### *A. Fuzzy C-Means*:

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. FCM [5] is a associate algorithm of fuzzy clustering which is based on K-means concepts to partition dataset directed toward clusters. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula:

$$\mu_{ij}=1/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(2/m-1)}$$

$$v_j=(\sum_{i=1}^{n}(\mu_{ij})^m x_i)/(x\sum_{i=1}^{n}(\mu_{ij})^m)$$

where, 'n' is the number of data points. 'vj' represents the jth cluster center. 'm' is the fuzziness index m ∈ [1,∞]. 'c' represents the number of cluster center. 'µij' represents the membership of ith data to jth cluster center. 'dij' represents the Euclidean distance between ith data and jth cluster center. Main objective of fuzzy c-means algorithm is to minimize:

$$J(U,V)=\sum_{i=1}^{n}\sum_{i=1}^{c}(\mu_{ij})^m \| x_i-y_j \|^2$$

where, '||xi – vj||' is the Euclidean distance between ith data and jth cluster center.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

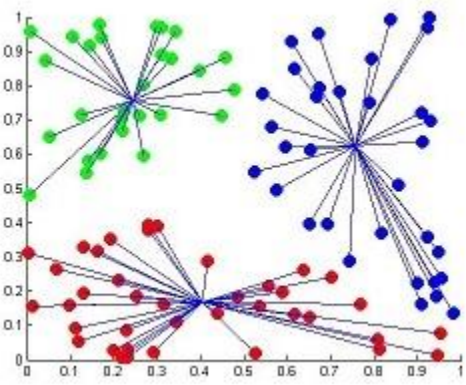

*Fig I: Result of Fuzzy c-means clustering*

### A. *BIRCH*:

BIRCH algorithm [6] builds a dendrogram known as a clustering feature tree (CF tree). The CF tree can be built by scanning the dataset in an incremental and dynamic way. Thus, it does not wish the complete dataset in advance. It has two main phases: the database is first scanned to build an inmemory tree, and then the algorithm is applied to cluster the leaf nodes. CF-tree is a height-balanced tree which is based on two parameters: branching factor B and threshold T. The CF tree is built while scanning the data. When a data point is encountered, the CF tree is traversed, starting from the root and choosing the closest node at each level. If the closest leaf cluster for the current data point is finally identified, a test is performed to see whether the data point belongs to the candidate cluster or not. If not, a new cluster is created with a diameter greater than the supposing T. BIRCH can truly find a useful clustering mutually a single scan of the dataset and refresh the quality further with a few additional scans. It can also handle noise effectively. However, BIRCH commits not function well when clusters are not spherical now it uses the work of extension or diameter to act the boundary of a cluster. In addition, it is order-sensitive and may generate different clusters for different orders of the same input data.

### B. *DENCLUE:*

Density-based Clustering (DENCLUE) uses an influence work to explain the impact of a point about its neighborhood while the overall density of the data space is the sum of influence functions from all data. The DENCLUE algorithm [7] analytically models the cluster distribution according to the sum of in_uence functions of all of the data points. Clusters are determined using density attractors, local maxima of the overall density function. To compute the sum of influence functions a grid structure is used. DENCLUE scale swell (O(N)), can find arbitrary-shaped clusters, is noise resistant, is in sensitive to the data ordering, but suffers from its sensitivity to the input parameters. The curse of dimensionality phenomenon heavily affects Denclue's effectiveness. Moreover, similar to hierarchical and partitioning techniques, the output, e.g. labeled points with cluster identifier, of density-based methods cannot be easily assimilated by humans. Advantages are Discovery of arbitrary-shaped clusters with varying size and Resistance to noise and outliers.

### C. *OPTIMAL GRID:*

OptiGrid [8] to address several aspects of the "curse of dimensionality": noise, scalability of the grid construction and selecting complementary attributes by optimizing the density field round the data space. OptiGrid uses density estimations to show the centers of clusters as the clustering was done for the DENCLUE algorithm [9]. A cluster is a region of concentrated density centered on a strong density attractor or local maximum of the density function with density above the noise threshold. Clusters may also have multiple centers if the centers are strong density attractors and there exists a way between them above the noise threshold. By recursively partitioning the feature space into multidimensional grids, OptiGrid creates an optimal grid-partition by constructing the best cutting hyperplanes of the space. These cutting planes cut the space in areas of low density (i.e. local minima of the density func- tion) and preserve areas of an arm and a leg density or clusters, specifically the cluster centers (i.e. local maxima of the density function). The cutting hyperplanes are found using a set of contract- ing linear projections of the feature space. The contracting projections create upper bounds for the density of the planes orthogonal to them. Namely, for any point x, in a contracting projection P, then for any point y one that P (y) = x, the density of y is at most the density of x.

### D. *EXPECTATION-MAXIMIZATION (EM)*

EM algorithm iteratively approximates the unknown model parameters with two steps: the E step and the M step. EM algorithm [10] is designed to add the

maximum likelihood parameters of a statistical model in multiple situations, such as the, such where the equations cannot be solved directly. The EM algorithm is guaranteed to find a local maximum for the model parameters estimate. In the E step (expectation), the current model parameter values are used to evaluate the posterior distribution of the latent variables. Then the objects are fractionally assigned to each cluster based on this posterior distribution. In the M step (maximization), the fractional assignment is given by re-estimating the model parameters with the maximum likelihood rule. The major disadvantages for EM algorithm are: the requirement of a non-singular covariance matrix, the sensitivity to the selection of initial parameters, the possibility of convergent evolution to a local optimum, and the slow convergence rate. Moreover, there would be a decreased precision of the EM algorithm within a finite number of steps [11].

## VI. CONCLUSION

This survey provided a comprehensive study of the clustering algorithms proposed in the literature. In order to reveal future directions for developing new algorithms and to guide the assignment of algorithms for big data, we proposed a categorizing framework to classify a number of clustering algorithms. The categorizing context is developed from a theoretical aspect that would automatically recommend the closely suitable algorithm(s) to network experts while hiding all technical details irrelevant to an application. Thus, even future clustering algorithms could be incorporated into the framework according to the proposed criteria and properties. Furthermore, the closely representative clustering algorithms of each segment have been empirically analyzed everywhere a vast number of evaluation metrics and traffic datasets.

## REFERENCES

1) Abbasi and M. Younis, ``A survey on clustering algorithms for wireless sensor networks," Comput. Commun., vol. 0, nos. 14_15, pp. 2826_2841, Oct. 2007.

2) C. C. Aggarwal and C. Zhai, ``A survey of text clustering algorithms," in Mining Text Data. New York, NY, USA: Springer-Verlag, 2012, pp. 77_128.

3) J. Brank, M. Grobelnik, and D. Mladeni¢, ``Asurvey of ontology evaluation techniques," in Proc. Conf. Data Mining DataWarehouses (SiKDD), 2005.

4) R. Xu and D. Wunsch, ``Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645_678, May 2005.

5) J. C. Bezdek, R. Ehrlich, andW. Full, ``FCM: The fuzzy c-means clustering algorithm," Comput. Geosci., vol. 10, nos. 2_3, pp. 191_203, 1984.

6) T. Zhang, R. Ramakrishnan, and M. Livny, ``BIRCH: An ef_cient data clustering method for very large databases," in Proc. ACM SIGMOD Rec., Jun. 1996, vol. 25, no. 2, pp. 103_114.

7) A. Hinneburg and D. A. Keim, ``An ef_cient approach to clustering in large multimedia databases with noise," in Proc. ACM SIGKDD Conf. Knowl. Discovery Ad Data Mining (KDD), 1998, pp. 58_65.

8) Alexander Hinneburg, Er Hinneburg, and Daniel A. Keim. An efficient approach to clustering in large multimedia databases with noise. In Proc. 4rd Int. Conf. on Knowl- edge Discovery and Data Mining, pages 58–65. AAAI Press, 1998.

9) Alexander Hinneburg and Daniel A. Keim. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In VLDB'99, pages 506–517. Morgan Kaufmann, 1999

10) A. P. Dempster, N. M. Laird, and D. B. Rubin, ``Maximum likelihood from incomplete data via the em algorithm," J. Roy. Statist. Soc., Ser. B, vol. 39, no. 1, pp. 1_38, 1977.

11) M. Meil and D. Heckerman, ``An experimental comparison of several clustering and initialization methods," in Proc. 14th Conf. Uncertainty Artif. Intell. (UAI), 1998, pp. 386_395.

12) J. Brank, M. Grobelnik, and D. Mladeni¢, ``Asurvey of ontology evaluation techniques," in Proc. Conf. Data Mining DataWarehouses (SiKDD), 2005.

13) A. P. Dempster, N. M. Laird, and D. B. Rubin, ``Maximum likelihood from incomplete data via the em algorithm," J. Roy. Statist. Soc., Ser. B, vol. 39, no. 1, pp. 1_38, 1977.

14) S. Guha, R. Rastogi, and K. Shim, ``Cure: An ef_cient clustering algorithm for large databases," in Proc. ACMSIGMOD Rec., Jun. 1998, vol. 27, no. 2, pp. 73_84.

15) S. Guha, R. Rastogi, and K. Shim, ``Rock: A robust clustering algorithm for categorical attributes," Inform. Syst., vol. 25, no. 5, pp. 345_366, 2000.

16) J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2006.

17) A. Hinneburg and D. A. Keim, ``An ef_cient approach to clustering in large multimedia databases with noise," in Proc. ACM SIGKDD Conf. Knowl. Discovery Ad Data Mining (KDD), 1998, pp. 58_65.

18) Hinneburg and D. A. Keim, ``Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering," in Proc. 25th Int. Conf. Very Large Data Bases (VLDB), 1999, pp. 506_517.

19) Z. Huang, ``A fast clustering algorithm to cluster very large categorical data sets in data mining," in Proc. SIGMODWorkshop Res. Issues Data Mining Knowl. Discovery, 1997, pp. 1_8.

20) L. Hubert and P. Arabie, ``Comparing partitions," J. Classi_cation, vol. 2, no. 1, pp. 193_218, 1985.

21) A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Upper Saddle River, NJ, USA: Prentice-Hall, 1988.

22) G. Karypis, E.-H. Han, andV.Kumar, ``Chameleon: Hierarchical clustering using dynamic modelling," IEEE Comput., vol. 32, no. 8, pp. 68_75, Aug. 1999.

23) L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. New York, NY, USA: Wiley, 2009.

24) T. Kohonen, ``The self-organizing map," Neurocomputing, vol. 21, no. 1, pp. 1_6, 1998.

25) J. MacQueen, ``Some methods for classi_cation and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Statist. Probab. Berkeley, CA, USA, 1967, pp. 281_297.

26) M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, ``A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. ACM SIGKDD Conf. Knowl. Discovery Ad Data Mining (KDD), 1996, pp. 226_231.

27) A. Fahad, Z. Tari, A. Almalawi, A. Goscinski, I. Khalil, and A. Mahmood, ``PPFSCADA: Privacy preserving framework for SCADA data publishing," Future Generat. Comput. Syst., vol. 37, pp. 496_511, Jul. 2014.

28) Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, ``Toward an ef_cient and scalable feature selection approach for internet traf_c classi _cation," Comput. Netw., vol. 57, no. 9, pp. 2040_2057, Jun. 2013.

29) H. Fisher, ``Knowledge acquisition via incremental conceptual clustering," Mach. Learn., vol. 2, no. 2, pp. 139_172, Sep. 1987.

30) J. H. Gennari, P. Langley, and D. Fisher, ``Models of incremental concept formation," Artif. Intell., vol. 40, nos. 1_3, pp. 11_61, Sep. 1989.