Estimation of Optimal Number of Clusters: A New Approach to Minimizing Intra-Cluster Communication Cost in WSNS

Emmanuel Effah , Ousmane Thiare

Abstract: ***Clustering of sensor nodes (SNs) is an unsurpassed energy management method in wireless sensor networks (WSNs) that ensures efficient energy balancing and duty-cycling, and improves the lifespan of the network by minimizing intra-cluster communication cost. Thus, since any incidences of misclustering shortens the lifespan of WSN, this paper presents an efficient, unbiased and more stable approach for evaluating the optimality of event-reporting (E-R) clusters in WSNs using the theory symbolic classifiers. Using realistic dataset derived from 1500 randomly deployed SNs, our results showed that the optimal number of clusters that guarantee optimal E-R accuracy and lengthened WSN lifespan by minimizing the intra-cluster communication costs are 240 clusters for classical K-Means method and 390 clusters for Extreme Learning Machine-Auto Encoder (ELM-AE). This method outperformed the classical inertia-based approach by establishing the optimal proxy E-R clusters which ensures higher E-R accuracy and energy efficiency of SNs. The experiment was done using realistic dataset extracted from randomly deployed 1500 SNs, and so our result is credible for the assessment of cluster qualities in other WSNs.***

Index Terms**: Wireless Sensor Networks,** ***Intra-Cluster Communication Costs, Recall and Precision***

I. INTRODUCTION

Power management has been a key research drive in wireless sensor networks (WSNs), and the efficient duty-cycling of the radio transceiver, the main power consuming unit [1] of the SN, seems to be the most reliable answer. In addition, it has been well established that the longer the distance of communication, the higher the energy consumed by the SNs and vice versa. Thus, the energy consumed by a SN to transmit a packet is equal to d^k, where d the distance of data transmission and k is an interval constant. Clustering of SNs is a superb way of addressing this challenge because clustering reduces the amount of data traffic most especially to the base station as well as the communication distances of SNs. However, the confirmatory optimality metrics of the clusters that can be used to ascertain the actual quality of the proxy E-R cluster remain undiscovered in WSNs, and the present study fills that gap.

In clustering paradigm, SNs are grouped to enhance local processing and efficient data transmission to the Sink. Every cluster has a cluster head (CH) and or a gateway node. For a more scalable and energy-balanced WSNs, clustering is the most reliable approach [1]. Selection of CH is very crucial and so deriving an efficient algorithm to unravel the most accurate metrics for higher clustering efficiencies has been well researched [2–10]. Fig. 1 presents the author - CH selection metrics as proposed in their various algorithms. Among principal metrics introduced by this study is the average relative distance of every SN from the present event source(s).

The principal aim of clustering sensor nodes (SN) is to maximize the WSN's lifespan through efficient energy management (by minimizing the cost of intra-cluster communication [2,9]). A critical synthesis the metrics in Fig. 1 reveals that, the main objective of the countless researches on the efficient CH selection metrics is imbued in efficient energy management of the SN which can be effectively achieved by minimizing intra-cluster communication cost. This implies that any incidence of misclustering is detrimental to our intent of maximizing the lifespan of the WSN through efficient energy management. Since there is no perfect clustering algorithm yet, the need for efficient and unbiased approach for evaluating the quality of the formed clusters is urgently desired.

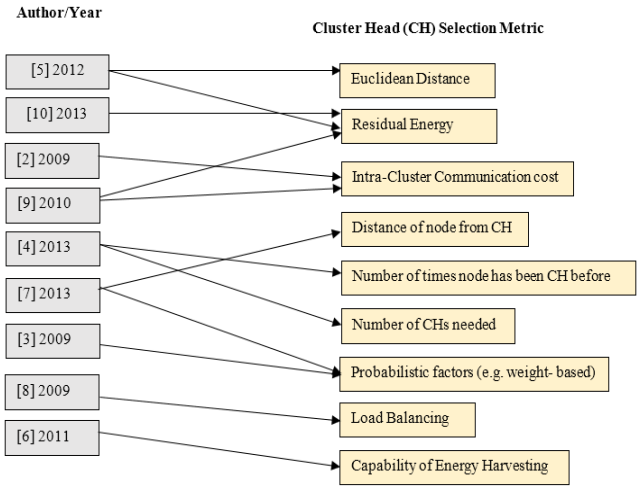

Fig.1. Mapping: CH Selection Metrics

From Table 1, every clustering or classification algorithm obeys this confusion matrix table. Thus,

Revised Manuscript Received on May 28, 2019.

Emmanuel Effah, Department of Computer Science, University Gaston Berger, St. Louis, Senegal.
Ousmane Thiare, Department of Computer Science, University Gaston Berger, St. Louis, Senegal.

irrespective of the clustering algorithm and the principal components or metrics
used, the tendency of misclustering is unavoidable.

Table 1. Confusion Matrix

	Predicted (i)	Predicted (j)
Actual Cluster (i)	TN (C_{ii})	FP (C_{ij})
Actual Cluster (j)	FN (C_{ji})	TP (C_{jj})

where:
(Cjj) = TP: Actual: positive and Prediction: positive.
(Cji) = FN: Actual: positive but Prediction: negative.
(Cii) = TN: Actual: negative and Prediction: negative.
(Cij) = FP: Actual: negative but Prediction: positive.

II. LITERATURE REVIEW

Classification methods such as clustering have become required approaches for the manipulation and analysis of large data. Clustering or unsupervised classification technique establishes strong cohesion among similar clusters [11] and ensure distinction between different clusters without the need for predefined classes of the dataset.
The fundamental challenge here is how to ascertain that the resulting clusters are of the desired quality or form the optimal number of clusters. In WSN, discovering a cluster quality index would not only improve decision regarding clustering algorithm to be deployed to obtain the ideal number of clusters of SNs, but could also facilitate effective data transmission and lengthen the lifespan of the network by minimizing the intra-cluster communication cost. The current alternative clustering quality evaluation (CQE) approaches [12–15] are predominantly distance-based indexes with their basic concepts from homogeneity - heterogeneity properties of clusters or intra-cluster and inter-cluster inertia [15, 11]. These Inertia-based CQE procedures are method-dependent and do not give accurate estimations when handling complex data [16,11,17] as in the case of highly densed WSNs.

A. Intra-Cluster Inertia

This measures the degree of homogeneity between the SNs within a cluster. It calculates the distances of the SNs with reference to the CH in lieu of the profile of the cluster. It can be densed as [16]:

$$Intra - Cluster\ Inertia = \frac{1}{|C|}\sum_{c\in C}\frac{1}{|C|}\sum_{d\in c}||P_c - P_d||^2$$

where:
C: is the set of clusters formed, *d:* is the data associated with the cluster and p_x: is the profile vector of clustered element x.

B. Inter-Cluster Inertia

The inter-cluster inertia measures the degree of heterogeneity amongst the clusters [16]. It computes the distances between the CHs representing the profiles of the various clusters of the partition [16] using:

$$Inter - Cluster\ Inertia = \frac{1}{|C|^2 - |C|}\sum_{c\in C}\frac{1}{|C|}\sum_{c'\in C, c'=c}||P_c - P_{c'}||^2$$

Using these traditional CQEs, a clustering result is accurate if the intra-cluster distances are minimum as compared to its inter-cluster distances [16]. However, the authors in [13] established that these distance-based paradigms can be biased and depend strongly on the clustering algorithms used. Fault Management and Multichannel communications techniques in WSNs are predominantly reliant on SNs clustering architecture [18], [19], and the complexity extremes in recent WSNs' deployments demand complex clustering procedures to handle the resulting multidimensional dataset metrics.
There is therefore the need to build an optimal clustering model and this calls for a more resilient and accurate CQE metrics to validate these clustering algorithms. Since traditional indexes are unreliable in the determination and validation of an ideal clustering model from complex and multidimensional dataset, [20], there is a need for one especially in the field of WSNs.
The stipulated problems could be managed using Recall/Precision indexes (Information Retrieval - IR approach), symbolic-based classifiers, that uses post clustering data (properties or metrics) of each cluster without prior knowledge of clusters profiles [21]. The symbolic equivalence these CQEs approach uniquely makes it totally independent of clustering algorithms and mode of operation [11].
With the recent technological improvements in microelectronics, affordable and high processing capability-micro-sized SNs are available to be exploited in myriad application areas in our daily lives for optimum benefits [5]. Clustering technique is one of the best ways of dealing with power issues, fault tolerance and efficient data transmission in WSNs. This has made it crucial to build an optimal clustering model but before such a model can be built, there will a need for an efficient and unbiased cluster quality evaluator (CQE) to assess the clusters formed from the set of clustering methods. This study plays a pivotal role in the discovery of Optimal clustering model for WSNs.

III. RECALL(R), PRECISION(P) AND F-MEASURE (F) IR APPROACH

A. General CQE in View

In IR systems:

$$R = \frac{TP}{TP+FN} = \frac{C_{jj}}{C_{ji}+C_{jj}} \quad (1)$$

$$P = \frac{TP}{TP+FP} = \frac{C_{jj}}{C_{ij}+C_{jj}} \quad (2)$$

$$F - Measure = \frac{2(R*P)}{R+P} \quad (3)$$

Thus, R and P are inversely correlated supervised indexes and F presents the best compromise between R and P [11].

Using the principles illustrated in these equations, the ensuing content will present the clustering or unsupervised version of R and P indexes for cluster optimality evaluation using different clustering methods [22] from shared metrics/properties perspectives. Assume that the principal metrics of a cluster content of our data are averaged in the range [0,1].

Let *C* = cluster sets formed from clustering *N* SNs, and the local R and P indexes of any metric or property m of c (cluster) is given by the expressions:

$$R_c(m) = \frac{|c_m^*|}{|N_m^*|}, \quad P_c(m) = \frac{|c_m^*|}{|c|} \tag{4}$$

Where: X_m^* indicates set X restrictions imposed on its elements with the metric m.

To estimate the clustering quality of all the clusters, the averaged Macro-Recall (R_M) and Macro-Precision (P_M) [11,16] must be invoked as:

$$R_M = \frac{1}{|C|}\sum_{c\in C}\frac{1}{|S_c|}\sum_{m\in S_c} R_c(m) \quad [16] \tag{5}$$

$$P_M = \frac{1}{|C|}\sum_{c\in C}\frac{1}{|S_c|}\sum_{m\in S_c} P_c(m) \tag{6}$$

where S_c represents unique metric set of cluster c, and its definition is:

$$S_c = \{m \in d, d \in c | \overline{W_{c'}^m} = Max\ c' \in C(W_{c'}^m)\} \tag{7}$$

$\bar{C}$ represents the unique set of clusters extracted from the clusters of *C* that validates: $\bar{C} = \{c \in C | S_c \neq \emptyset\}$ and finally:

$$\overline{W_{c'}^m} = \frac{\sum_{d\in c} W_d^m}{\sum_{c'\in C}\sum_{d\in c'} W_d^m} \tag{8}$$

where W_x^m represents the weight of the metric *m* for element x.

As well established by Lamirel et al [20,21], if both values of R_M and P_M (Eq. 5 and Eq. 6) approach unity, then $\bar{C}$ (unique set of clusters) satisfies the conditions of Galois lattice. These measures help to assess the degree of how numerical clustering models equate Galois lattice classifier.

R_M and P_M indexes compute mean values of R and P for each cluster of the SNs and exhibit contrary traits based on the number of clusters, hence, they are cluster-based quality measures. Explicitly, the optimal number of clusters of any clustering algorithm with any associated dataset can be estimated using these indexes and the best clustering result will occur at the value which minimizes the difference between these two indexes.

However, they suffer similar defect as inertia-oriented indexes which is their inability to detect degenerated clustering results [11]. However, the metric-based supporting indexes of Micro − R(Rm) and Micro − P(Pm) which work on Recall/Precision mean values of the unique metrics without depending on the structure of clusters [11] correct this defect using the Eq.9 and Eq. 10. They deployed cumulative operation of Micro-Precision to address this defect.

$$R_m = \frac{1}{|L|}\sum_{c\in C, m\in S_c} R_c(m) \tag{9}$$

$$P_m = \frac{1}{|L|}\sum_{c\in C, m\in S_c} P_c(m) \tag{10}$$

where L represents the size of the data description space [16]. However, our dataset used was free from this defect and so this approach was not considered.

IV. Results and Discussion

The results were obtained by plotting number of clusters against the averaged principal contents of each CQE indexes as presented in Fig. 2 and Fig. 3. From Fig. 2, it is evident that inertia-based indexes produced unstable behaviour and consequently makes it an incredible estimator of optimal number of clusters in both K-Means and ELM-AE approaches. Inertia-based indexes also deny as the privilege of selecting an efficient clustering model as well as determining the optimal number of clusters.

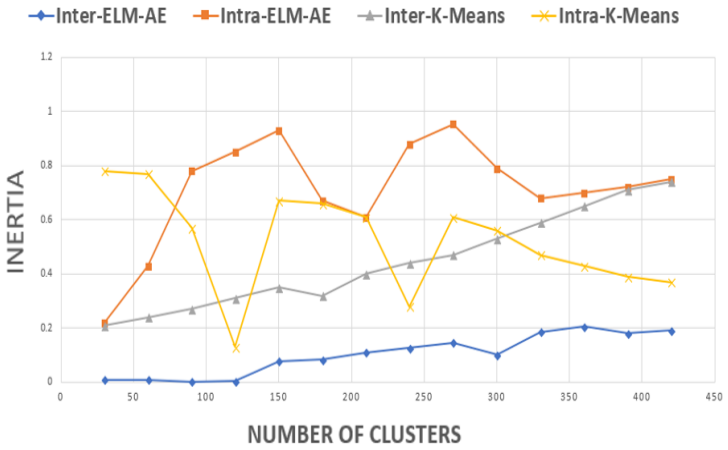

Fig.2. Intra and Inter Cluster Inertia Indexes (K-Means and ELM-AE) Against No. of Clusters

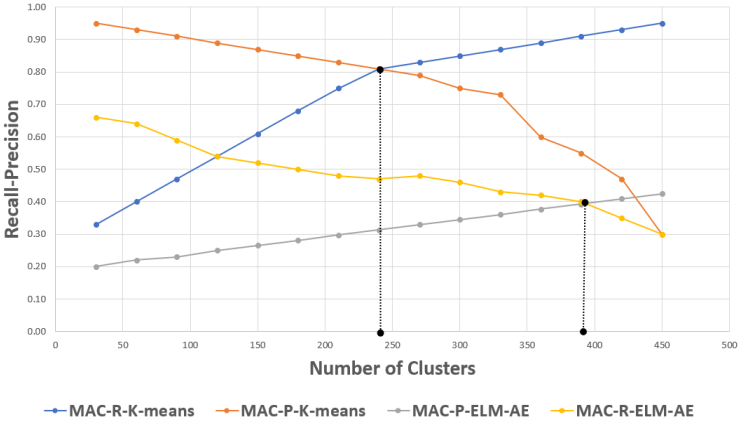

Fig.3. RM/ PM Indexes (K-Means and ELM-AE) Against No. of Clusters

However, Fig.3 illustrates the context of R_M/ P_M indexes as being more stable and hence credible enough for the estimation of the optimal number of clusters in

both methods. This optimal number of clusters occurs at the intersection between the R_M and P_M values. From Fig. 3, these optimal values are 240 and 390 clusters for K-Means and ELM-AE respectively. This defines the optimally minimum intra-cluster communication cost of the SNs. In other words, this point specifies the best clustering topology with the most minimum intra-cluster distances that will consequently minimize intra-cluster communication cost in order to maximize the lifespan of the WSN. Also, the fact that we obtained different optimal number of clusters using different clustering methods affirms the need for further studies into different clustering models to establish an optimal clustering model for WSNs. This approach presented in this paper gives more accurate assessment of the quality of the clusters formed in WSNs.

V. CONCLUSION

This study has established that the R_M/P_M indexes yields more stable, reliable, credible and optimal number of clusters of K-Means and ELM-AE methods required to maximize WSNs lifespan (reduced intra-cluster communication cost) as 240 and 390 respectively. Aside prolonging the lifespan of the SNs, best energy management techniques such as proposed here also prevents the occurrence of faults and improves fault tolerance because the root cause of faults in WSN is power mismanagement [24]. The uniqueness of our approach in comparison with distance-based methods for evaluating the quality clusters has been unveiled; the theoretical basis has also been established by exploiting links in symbolic classification. The experiment was done using realistic dataset generated from randomly deployed 1500 SNs, and this implies that our result is credible for the assessment cluster quality in WSNs.

REFERENCES

1. M. Geeta, “Various clustering techniques in wireless sensor network,” International Journal of Computer Applications Technology and Research, Vol. 3, Issue 6, ISSN: 23198656, pp. 381 – 384, 2014.
2. A. Bereketli,Ozgur B. Akan, “Event-to sink directed clustering in wireless sensor networks,” Next generation Wireless Communications Laboratory (NWCL) Department of Electrical and Electronics Engineering Middle East Technical University, Ankara ,Turkey,, pp. 06531 978–1–4244–2948–6/09, 2009.
3. Basilis Mamalis, Damianos Gavalas, Charalampos Konstantopoulos, and Grammati Pantziou, “Clustering in wireless sensor networks,” Zhang/RFID and Sensor Networks AU7777C012,pp.324—364,2009.
4. G.Y. Durga Devi, “Clustering algorithms in wireless sensor networks-a survey,” ISSN (Online): 2347-2820, Issue-2, vol. 1, pp. 1–9, 2013.
5. P. Sasikumar ,Sibaram Khara, “K-means clustering in wireless sensor networks,” 978-07695-4850-0/12 26.002012IEEEDOI10.1109/CICN.2012.136,pp.1 − −2,2012.
6. Pengfei Zhang, Gaoxi Xiao and Hwee-Pink Tan, “A preliminary study on lifetime maximization in clustered wireless sensor networks with energy harvesting nodes,” 978-14577-0031-6/11/26.00IEEE, pp.1−2, 2011.
7. Ravi Tandon, Biswanath Dey and Sukumar Nandi, “Weight based clustering in wireless sensor networks,” 978-1-4673-5952-8/13/31.00IEEE, pp.1—2,2013.
8. Shujuan Jin, Keqiu Li, “Lbcs: A load balanced clustering scheme in wireless sensor networks,” Third International Conference on Multimedia and Ubiquitous Engineering, pp. 1–2, 2009.
9. Tianqi Wang, Wendi Heinzelman, and Alireza Seyedi, “Maximization of data gathering in clustered wireless sensor networks,” 2010, pp. 1–2.
10. Tony Ducrocq, Nathalie Mitton, Michal Hauspie, “Energy-based clustering for wireless sensor network lifetime optimization,” in WCNC - Wireless Communications and Networking Conference - 2013, Apr 2013, Shanghai, China. 2013. ¡hal00767690.
11. Lamirel, J.-C, “Reliable clustering quality estimation from low to high dimensionnal data,” in 11th Workshop on Self-Organizing Maps WSOM 2016, Houston, TX, USA (2016).
12. A. D. Gordon, “External validation in cluster analysis,” in Bulletin of the International Statistical Institute, 51(2), 353-356 (1997), Response to comments. Bulletin of the International Statistical Institute 51(3), (1998), 414-415. 10.
13. M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” Journal of Intelligent Information Systems, 17:2/3, pp. 147155, 2001.
14. Lamirel, J.-C., Cuxac P., Chivukula A.S., Hajlaoui K., “Optimizing text classication through efcient feature selection based on quality metric,” Journal of Intelligent Information Systems, Special issue on PAKDDQIMIE 2013, pp. 1–18, 2014.
15. L. Lebart, A. Morineau, and J.P. Fenelon, “Traitement des donnees statistiques,” Dunod, Paris, 1979.
16. Jean-Charles Lamirel, “A new efficient and unbiased approach for clustering quality evaluation,” 2012.
17. Kolesnikov, A. and Trichina, E. and Kauranne, T., “Estimating the number of clusters in a numerical data set via quantization error modeling,” Pattern Recognition, 48(3), p. 941952, 2015.
18. E. Ahmed, A. Gani , S. Abolfazli, L. J. Yao, and S. U. Khan, “Multichannel and cognitive radio approaches for wireless sensor networks,” Communications Surveys, vol.18, pp. 795–823, 2014.
19. M. Boban, A. Festag, “Service actuated multichannel operation for vehicular communication,” computer communications, pp. 17–26, 2016.
20. R. Kassab, and J.-C. Lamirel, “Feature based cluster validation for high dimensional data,” IASTED International Conference on Articial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008, pp. 97–103, 2008.
21. J-C. Lamirel, S. Al-Shehabi, C. Francois, and M. Hofmann, “New classication quality estimators for analysis of documentary information: application to patent analysis and web mapping,” Scientometrics, pp. 445–562, 2004.
22. J.-C. Lamirel, S. Al Shehabi, “Multisom: A multiview neural model for accurately analyzing and mining complex data”, fourth international conference on coordinated multiple views in exploratory visualization (cmv 06),” 2006.
23. S. Al Shehabi, “Multisom: A multiview neural model for accurately analyzing and mining complex data”, fourth international conference on coordinated multiple views in exploratory visualization (cmv 06),” 2006.
24. E. Effah, and O. Thiare, “Survey: Faults, fault detection and fault tolerance techniques in wireless sensor networks,” International Journal of Computer Science and Information Security, IJCSIS), pp. 1–14, Vol. 16, No. 10, ISSN 1947-5500, October 2018.

AUTHORS PROFILE

Emmanuel Effah is pursuing his PhD in Computer Science at University Gaston Berger, Senegal. He is a Lecturer (on study leave) at the Department of Computer Science and Engineering, University of Mines and Technology, Ghana. His research interests include wireless sensor networks (WSNs), smart grid systems, data mining and smart systems' technology. He has authored many research papers published in reputable journals and conference proceedings.

Ousmane Thiare received his PhD in Computer Science at University of Cergy-Pontoise, Paris-France. He is Vice Chancellor of University Gaston Berger, Saint Louis, Senegal. His research interests are Internet of Things (IoT), Wireless Sensor Networks, Mobile Ad hoc Networks, Parallel and Distributed Systems, Peer-to-Peer Systems, Routing and Data replication and Grid and Cloud Computing. He has published research papers at national and international journals, conference proceedings as well as chapters of books.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication