Management Science Letters 3 (2013) 967–972

Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

IP2P K-means: an efficient method for data clustering on sensor networks

Peyman Mirhadi[*a], Sajjad Zandinia[a], Azadeh Goodarzipour[a], Siamak Salimi[b] and Hossein Goodarzipour[a]

[a]*Farnas Aria Co., Zafar St., Tehran, Iran*
[b]*Qazvin Islamic Azad University, Qazvin, Iran*

C H R O N I C L E

Article history:
Received October 21, 2012
Received in revised format
5 January 2013
Accepted 18 January 2013
Available online
January 20 2013

Keywords:
Clustering algorithm
Wireless sensor network
Data stream
Network Simulator
Data aggregation

A B S T R A C T

Many wireless sensor network applications require data gathering as the most important parts of their operations. There are increasing demands for innovative methods to improve energy efficiency and to prolong the network lifetime. Clustering is considered as an efficient topology control methods in wireless sensor networks, which can increase network scalability and lifetime. This paper presents a method, IP2P K-means – Improved P2P K-means, which uses efficient leveling in clustering approach, reduces false labeling and restricts the necessary communication among various sensors, which obviously saves more energy. The proposed method is examined in Network Simulator Ver.2 (NS2) and the preliminary results show that the algorithm works effectively and relatively more precisely.

1. Introduction

During the past few years, there has been growing popularity among world's nation to use wireless communication devices, which has also created more interests in communication infrastructure caused emergence of wireless sensor networks (WSN) (ChitraDevia et al. 2012). These networks normally include intelligent sensors, which are equipped with some other advanced microsensors to detect their environment, a small processor or even a low range wireless communication device.

In such networks, sensors with communication together, make a global framework from the environment. In many sensor network usages, real time data processing and global meaningful techniques for intelligent and rapid decision makings are unavoidable (Khalil & Attea 2011; Schaffer et al. 2012). To take advantage of these models we need data mining on some information and the

*Corresponding author.
E-mail addresses: peyman.mirhadi@gmail.com (P. Mirhadi)

doi: 10.5267/j.msl.2013.01.021

primary concern is on how to cluster the data through an appropriate data mining technique to process a group of similar objects with common attributes. With sensor's data clustering, it is possible to get an overall wisdom to the manner of data distribution and clustering is the first step for processing the data (Aioffi, Valle et al. 2011). Clustering is also considered as one of the effective solutions to enhance energy efficiency and scalability of large-scale wireless sensor networks. The primary objective of clustering is to identify a subset of nodes in a wireless sensor network where all other nodes communicate with the network sink via these selected nodes (Bhardwaj, SoniDinesh et al. 2012). However, many existing clustering algorithms are tightly coupled with exact sensor locations derived through either triangulation techniques or extra hardware such as GPS equipment. However, in practice, it is difficult to detect sensor location coordinates precisely because there are different influencing factors such as random deployment, low-power and low-cost sensing devices (Ribas, Colonna et al. 2012; Silva, Chiky et al. 2012; Wei, Chen et al. 2012).

Since the nature of distributed and restricted network and communication resources is somehow unknown, it is necessary to make use of distributed algorithms. In this paper, we present a new distributed data clustering for sensor networks in terms of bandwidth, energy and memory restrictions (Liu & Li, 2012).

1.2. Sensor networks

Sensor networks are always dealt with a variety of challenges including energy, data processing, communication and routing restrictions. Design of protocols and routing algorithms in sensor networks to minimize the energy consumption is an area of open research. Routing protocols must include three main capabilities in networks: identification of topology changes, communication establishment in networks and detecting appropriate routes. In case of sleep state, existence of middle nodes increases packet transmission delay (Akkaya & Senel, 2009; Bajaber & Awan 2011).

2. Material and methods

In this paper, we present a clustering algorithm where the primary part of it is associated with data streaming processing and the other part is responsible of final data clustering. Because data stream is a continuous flow, data stream processing section of the algorithm is always on running stage. Therefore, it is impossible to store all data stream to main memory and so the proposed algorithm is approximate algorithms. The method tries to propose a solution where the target function is a constant approximate of efficient state of goal function.

The proposed algorithm uses location reduction for data stream process in restricted memory. Location reduction is the transformation of m data point to l ($l<m$), so l points contain characteristics of m points.

The proposed algorithm steps are:

1. Continue sampling as long as the majority of observed nodes, majority of gathered data by sensor node, have not exceeded the memory constraint (m).

2. Using classical k-mean algorithm, calculate O(k) of center for m point and replace m points. Use $2k$ central points, the location reduction process. The clustering is more precise if k is higher but the consumption memory is also higher. Consider a weight for every center. This weight is the points assigned to it.

3. Repeat step 1 and 2 until $m^2/2k$ point is read and m central point is obtained. These primer centers are considered level-1 centers.

4. Use k-means algorithm to reduce m level-1 centers to $2k$ level-2 centers.

5. keep the most m i-level center in memory and produce $2k$ $i+1\_level$ center if majority of i_level centers reaches m. Weight of new center is sum of weight of centers assigned to it.

6. If global clustering is obtained, apply *k-means* algorithm to all centers created to all levels, otherwise go to the previous step.

3. Results

To evaluate the efficiency of the proposed algorithm, two indicators are used. To measure the precise of proposed clustering, first indicator (LRI) shows percentage of errors on data points labeling. This indicator demonstrates the proportion of data whose cluster labels in two executions are different (distributed and non-distributed (central) algorithm), and is defined as follows,

$$LRI = ILC/n \times 100\%. \quad (1)$$

ILC is sum of points whose cluster labels are different in distributed and central algorithm and *n* is number of total points. Second indicator is the average distance between cluster centers in central and distributed approaches. We show this indicator with DRC_D:

$$DRC_D(j) = 1/p \left[\sum \| C_D^i(j) - C_j^c \| / \| C_j^c \| \right] \times 100\% \ , \qquad (j = 1, \ldots, k) \quad (2)$$

where *J* is the number of cluster, *p* is the number of network nodes, $C_D^i(j)$ is center of j^{th} cluster in i^{th} node in distributed algorithm, C_j^C is the center of j^{th} cluster in *k*-means central algorithm and $\| \ \|$ is second order norm.

$$\|X\| = \sqrt{x_1^2 + \cdots + x_1^2} \ . \quad (3)$$

To evaluate the performance of the proposed algorithm, a 400 seconds scenario with 24 nodes is run on *NS2*. To have different traffic, 8 UDP agents that have FTP application attached and 16 TCP nodes, which have CBR on them are consisted. Data rate for CBR are 15 Mbit/s and 11 Mbit/s for FTP. The proposed algorithm is run on every node Fig. 1. We have analyzed the results based on γ.

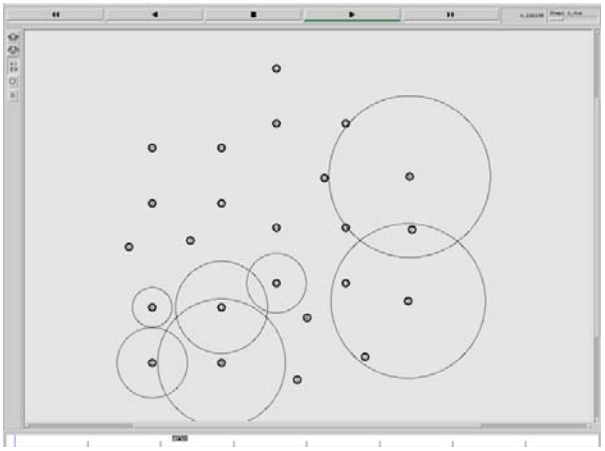

Fig. 1. Simulation environment in Network Simulator ver. 2

3.1. γ parameter

The proposed model of this paper uses γ parameter as defined by Bandyopadhyay and Giannella (2006) where termination provision is based on γ. This parameter is a criterion on testing center changes in two repetitions. It is obvious that lower value for γ means more clustering procession but

higher number of iterations will be required and it be more costly. Hence, it is important to select an appropriate amount of this parameter, which requires a tradeoff between clustering precision and communication cost. The changes of ILR in accordance with γ are shown on Fig. 2. As we can observe, with an decrease in γ, partial communication cost will increase and false labeling will be reduced.

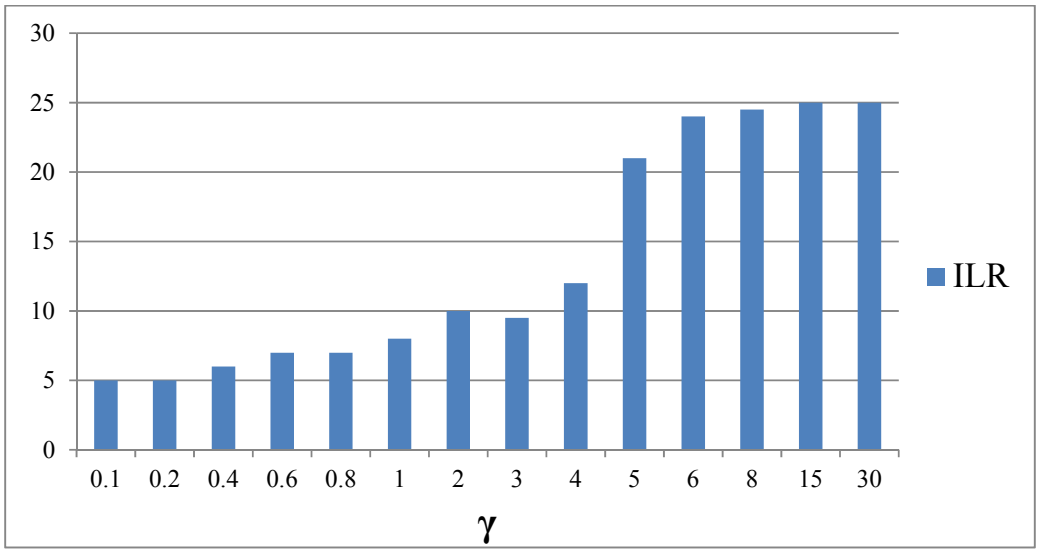

Fig. 2. The changes of ILR in accordance with γ

In Fig. 3, clustering is obtained from central *k*-means method (Modha & Spangler 2003) is shown. In Fig. 4 clustering obtained from proposed algorithm is shown. The similarity of these two figures approves the high precision of the proposed algorithm.

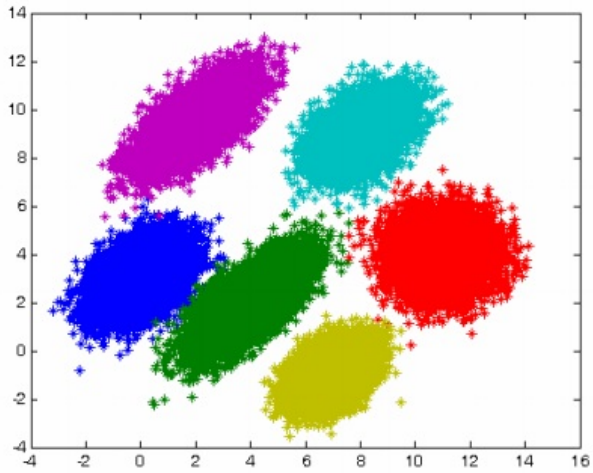

Fig. 3. Clustering obtained from central K-means algorithm (Modha & Spangler 2003)

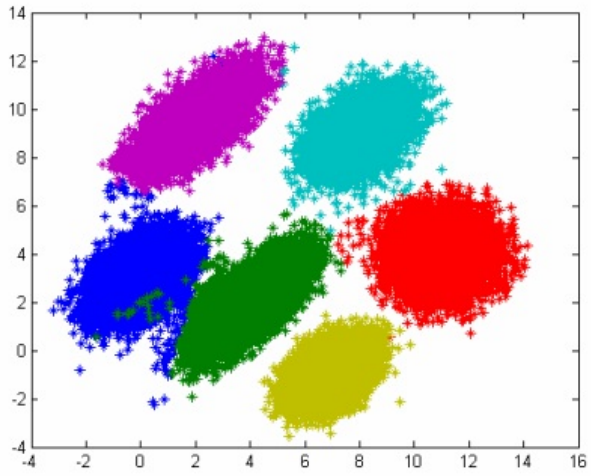

Fig. 4. Clustering obtained from IP2P K-means algorithm

4. Conclusion

Recent advancement in wireless communications and electronics has enabled the development of low-cost sensor networks. The sensor networks can be implemented in different applications and there are various technical issues where researchers are currently doing research on. A high-density wireless sensor network can be deployed for specific information-gathering. In such a network, sensors need to route their sensed data to a base station, consuming highly-limited and unreplenishable energy resource. Therefore, one of the most important issues in designing sensor data gathering algorithms is to minimize the energy consumption for network longevity while meeting certain requirements given, such as delay constraints, which may vary depending on specific applications or environmental situations.

In this paper, a new distributed data clustering on sensor networks was proposed where communication was the main reason of energy consumption in sensor networks. Therefore, the proposed algorithm attempted to reduce the communication and message interchange by trying to stop false labeling to save energy. The proposed algorithm was tested on a scenario with NS2 (Network Simulator ver.2) and the results showed efficient performance of the algorithm.

References

Aioffi, W.M., Valle, C.A., Mateus, G.R., & da Cunha, A. S. (2011). Balancing message delivery latency and network lifetime through an integrated model for clustering and routing in Wireless Sensor Networks. *Computer Networks* 55(13), 2803-2820.

Akkaya, K., Senel, F., & McLaughlan, B. (2009). Clustering of wireless sensor and actor networks based on sensor distribution and connectivity. *Journal of Parallel and Distributed Computing*, 69(6), 573-587.

Bajaber, F., & Awan, I. (2011). Adaptive decentralized re-clustering protocol for wireless sensor networks. *Journal of Computer and System Sciences,* 77(2), 282-292.

Bandyopadhyay, S., Giannella, C., Maulik, U., Kargupta, H., Liu, K., & Datta, S. (2006). Clustering distributed data streams in peer-to-peer environments. *Information Sciences*, 176(14), 1952-1985.

Bhardwaj, M., Soni, S., & Kotary, D.K. (2012). Comparative Analysis of Energy Efficient Routing Protocol for Wireless Sensor Network. *International Journal of Computer Applications*, 1, 65-59.

ChitraDevia, N., V. Palanisamy, et al. (2012). A Novel Distance for Clustering to Support Mixed Data Attributes and Promote Data Reliability and Network Lifetime in Large Scale Wireless Sensor Networks. *International Conference on Communication Technology and System Design 2011 - Procedia Engineering*.

Khalil, E. A., & Attea, B.A. (2011). Energy-aware evolutionary routing protocol for dynamic clustering of wireless sensor networks. *Swarm and Evolutionary Computation* 1(4), 195-203.

Liu, T., & Li, Q. (2012). An energy-balancing clustering approach for gradient-based routing in wireless sensor networks. *Computer Communications,* 35(17), 2150-2161.

Modha, D. S., & Spangler, W.S. (2003). Feature Weighting in k-Means Clustering. *Machine Learning* 52(3), 217-237.

Ribas, A.D., Colonna, J.G., Figueiredo, C.M.S., & Nakamura, E.F. (2012). Similarity clustering for data fusion in Wireless Sensor Networks using k-means. *The 2012 International Joint Conference on Neural Networks (IJCNN)*. Communication, Networking & Broadcasting ; Components, Circuits, Devices & Systems ; Computing & Processing (Hardware/Software) ; Engineered Materials, Dielectrics & Plasmas ; Fields, Waves & Electromagnetics ; Robotics & Control Systems, 1-7, doi: 10.1109/IJCNN.2012.6252430

Schaffer, P., Farkas, K., Horváth, A., Holczer, T. & Buttyán, L. (2012). Secure and reliable clustering in wireless sensor networks: A critical survey. *Computer Networks,* 56(11), 2726-2741.

Da Silva, A., Chiky, R., & Hébrail, G. (2012). A clustering approach for sampling data streams in sensor networks. *Knowledge and Information Systems* 32(1), 1-23.

Wei, H., Chen, L., & Zhang, Y. (2012). Expected number of Cluster Members clustering algorithm in wireless sensor networks. *Systems and Informatics (ICSAI), 2012 International Conference on* Communication, Networking & Broadcasting ; Components, Circuits, Devices & Systems ; Computing & Processing (Hardware/Software) ; Power, Energy, & Industry Applications, 1381 - 1384. DOI: 10.1109/ICSAI.2012.6223293