Energy Efficient Ensemble K-means and SVM for Wireless Sensor Network

Manal Abdullah , Ahlam Al-Thobaity, Afnan Bawazir , Nouf Al-harbe

Faculty of Computing and Information Technology , FCIT

King Abdulaziz University , KAU

maaabdullah@kau.edu.sa , Ahlamsaud2012@yahoo.com ,

Afnan-bawazir@hotmail.com, nouf_3954@yahoo.com

ABSTRACT

A wireless sensor network (WSN) consists of a large number of small sensors with limited energy. For many WSN applications, prolonged network lifetime is important requirements. There are different techniques have already been proposed to improve energy consumption rate such as clustering ,efficient routing , and data aggregation. In this paper, we present a novel technique using clustering .The different clustering algorithms also differ in their objectives. Sometimes Clustering suffers from more overlapping and redundancy data since sensor node's position is in a critical position does not know in which clustering it is belonging. One option is to assign these nodes to both clusters, which is equivalent to overlap of nodes and data redundancy occurs. This paper has proposed a new method to solve this problem and make use of the advantages of Support Vector Machine SVM to strengthen K-MEANS clustering algorithm and give us more accurate dissection boundary for each classes .The new algorithm is called K-SVM.Numerical experiments are carried out using Matlab to simulate sensor fields. Through comparing with classical K-MEANS clustering scheme we confirmed that K-SVM   algorithm has a better improvement in clustering accuracy in these networks.

Indexing terms/Keywords

WSN; clustering; k-means; SVM, network stability.

Academic Discipline And Sub-Disciplines

Computer Science / Machine Learning.;

SUBJECT CLASSIFICATION

patern recognition and classification

TYPE (METHOD/APPROACH)

K-mean / SVM

1.   INTRODUCTION

Recently many researchers have been conducted in wireless sensor networks WSN [1]. Due to their wide range of applications in the field of military surveillance, fire detection ,habitat monitoring , industry , health monitoring and many more. WSN is composed of a large number of randomly deployed sensor nodes. Sensor nodes are deployed within concerned area. WSN have at least one base station that works as a gateway between the sensor network and the outside world. Sensor nodes sense the phenomenon and send the data to the base station via single or multi-hop communication .Users access the data stored at the base station. Sensor nodes have limited battery power, limited memory and limited processing capabilities. So lifetime of WSN is limited by on-board energy of sensor nodes. Due to harsh deployed area, replacement or recharge of battery is not feasible. Lack of infrastructure and a large number of sensor nodes causes huge flow of message transfer through the network. As most of the energy is consumed during communication [2], the network life time has been a critical concern in WSN researches, and a number of research works [3] attempts to energy consumption and extend   network lifetime period by various techniques like routing, scheduling, aggregation, clustering.

In this paper, we aim to improve the network life time by using clustering. However, operation of the clustering algorithm is executed in rounds and each round is composed of two phases: setup phase and steady phase. Nodes are organized in independent sets or clusters. At least one cluster head is selected for each cluster. The sensed data is not directly sent to the base station but via respective cluster heads. Cluster head collects data of sensor nodes that belong to that cluster then sent it to the base station. This architecture is depicted in Fig 1. Some of clustering problems in WSNs is node overlap this arises when a node is adjacent to two clusters. One option is to assign these nodes to both clusters, which is equivalent to overlap of nodes and data redundancy. Some decisions algorithm are required to avoid this overlapping.
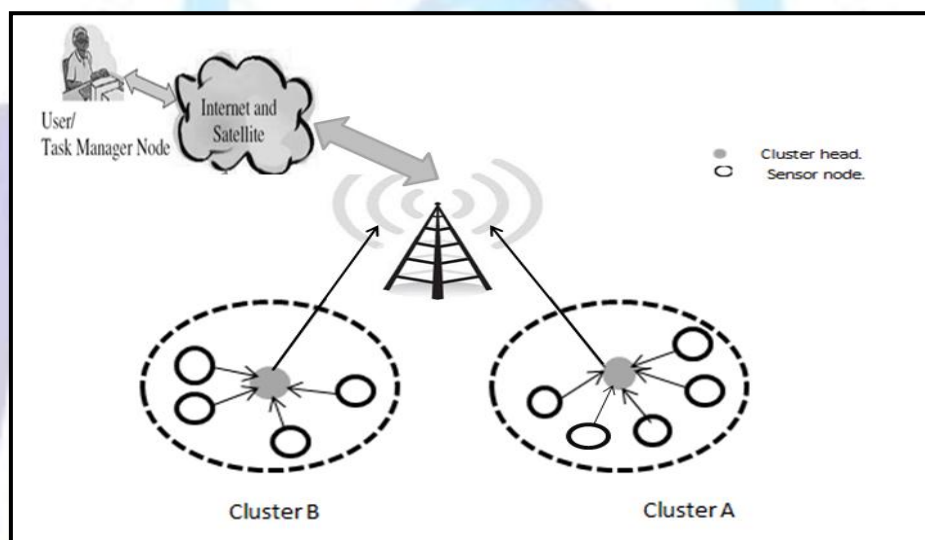

Figure1. Clustered Sensor Networks Architecture.

Main goal of this paper is to develop a novel collaborate between clustering in WSNs and the most informative patterns for the classification task which is SVM. In K-SVM we rely on k-means algorithm, which features simple, highly reliable, and fast-convergent iterations & re-clustering during failure states [4].To fuse this advantages with SVM whose advantages give a good dissection boundary for each classes [11]. We use Radial Base Function RBF as a kernel function to give a good estimate of clusters in a WSN.

The rest of this paper is organized as follow : related works are discussed in Section II. Section III discusses K-means algorithm in details. SVM for classifying in WSN and the difference between linear and nonlinear SVM are discussed in section IV. Section V presents the comparative results with analysis. Finally conclusions in section VI.

2. RELATED WORK

This section focuses on some clustering techniques have already been proposed for partitioning nodes in clusters. Some of the early and common clustering techniques are Distributed Clustering Algorithm (DCA) [5], and Spanning Tree (or BFS Tree) based Clustering [6]. In [7], Banerjee and Khuller suggested organizing the nodes in the network into hierarchical tier - structures. Within a particular cluster, data aggregation and forwarding are performed at cluster-head to reduce the amount of data transmitting to the base station. Cluster formation is usually based on energy of sensor nodes and sensor's proximity to cluster-head .Nodes other than cluster-head choose their cluster-head right after deployment and transmit sensed information to the cluster head. The role of cluster-head, being itself a sensor node, is to forward these data and its own data to the base station after performing data aggregation and forwarding. I n this study , the author did not address data redundancy.

Zdenek and Petr in [8] introduced an energy efficient scheduling for cluster-tree WSN. Clustering in this method uses Cluster-tree formation. Cyclic Scheduling for data transmission in Zigbee environment used Time-division Multiple Access (TDMA). The cluster is active only once during the schedule period which leads to so called cyclic behavior of periodic schedule when there are flow with opposite direction in a WSN. Message overhead or energy consumed is not addressed in this iterative method of clustering.

In [9] Jerzy suggested method of building clusters in the wireless sensor networks by using LS-SVM. With mixtures of kernels, the polynomial and the RBF kernel. Through comparing with classical clustering scheme, they stated that LS-SVM method has a better improvement in clustering accuracy and precision measures in these networks. Multiclass is not addressed in this method of clustering.

Exploit of the informative patterns of SVM for WSN node classification is the aim of our paper. K-SVM considers the data redundancy reduction that directly reflected on the energy consumption. K-SVM is a simple and energy efficient clustering process using K-means then SVM algorithm with radial base kernel function is applied to produce more accurate dissection boundary for each cluster, and this extends WSN network lifetime.

3. K-MEAN CLUSTERING ALGORITHM

We present the definition of the k-Mean Clustering in statistics and machine learning as an introduction to this paper, which k-means clustering is a method of cluster analysis that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

K-means algorithm is based on the Euclidian distances and cluster head choosing depends on maximum energy of nodes. The clustering algorithm (k-mean) , with each round, k nodes are selected randomly as centroids. Then the nodes group themselves to the nearest centroid. After the node has collected itself with the nearest centroid, then a new center is found and this center is the new centroid. This process maintain of repeating itself until all the nodes have been grouped into some cluster. Figure 2 illustrates the flow chart of K-means algorithm. Figure 2 illustrates the flow chart of K-means algorithm. Next shown in Figure 3 is the K-means algorithm that illustrates the four steps and the flow chart.
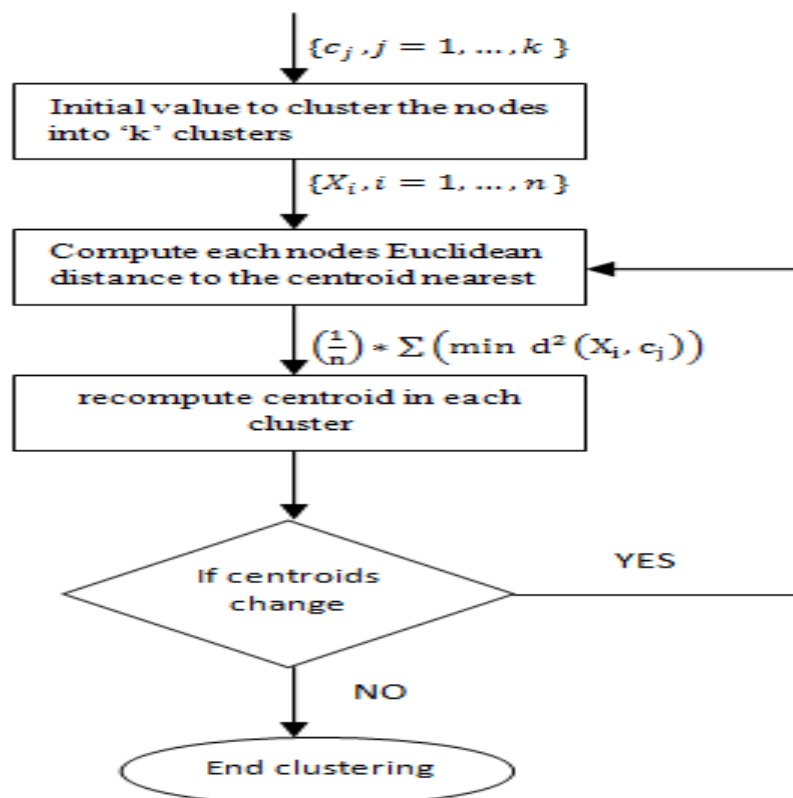
$$\{c_j , j = 1, \dots , k \}$$

Initial value to cluster the nodes into 'k' clusters

$$\{X_i , i = 1, \dots , n \}$$

Compute each nodes Euclidean distance to the centroid nearest

$$\left(\tfrac{1}{n}\right) * \sum \left(\min \ d^2 \left(X_i , c_j\right)\right)$$

recompute centroid in each cluster

If centroids change

YES

NO

End clustering

Figure 2 :Flow Chart of K-means Algorithm

Algorithm 1: K-Means Clustering

Input: $X = \{x_1, x_2, \ldots, x_n\}$ (set of nodes to be clustered)

k (number of clusters)

Output :

$C = \{c_1, c_2, \ldots, c_k\}$ (set of cluster centroids)

Initialization

for each $c_i \in C\ do$

$\quad \big| c_i \leftarrow x_j \in X$ (*random selection*)

end

repeat

$\qquad$ for each $x_i \in X\ do$

$\qquad minDist \leftarrow minDistance(x_i, c_j)\ j \in \{1 \ldots k\};$

$\qquad$ (based on Euclidean distance $\left(\frac{1}{n}\right) *$

$$\Sigma \left(\begin{array}{c} \min \\ j \end{array} d^2\,(X_i, c_j) \right),$$

$\qquad$ for i = 1 to n,)

$\qquad$ UpdateCluster(c_i)

$\quad end$

$until$ convergence

Figure 3: K-means Clustering Algorithm

The process of choosing cluster head starts after the centroid positions are finalized. We consider node highest energy in each cluster as Cluster Head "CH". If more than one node has same highest energy, the node nearest to the base station is selected as cluster head.

4. SVM FOR CLASSIFYING WSNS

SVM which refers to "Support Vector Machines" is a generalized machine learning algorithm that based on statistical learning theory. SVM algorithms have been successfully applied to minimum data redundancy in feature space. In this section, we exploit SVM method used to classify in WSNs "wireless sensor networks".

We present the Linear Support Vector Machine and Nonlinear Support Vector Machines. In Linear Support Vector Machine that can only solve problems that are linearly separable, i.e. where the node belonging to class 1 can be separated from the node belonging to class 2 by a hyper plane. While In many cases, a classification problem that is not linearly separable can be solved by applying Nonlinear Support Vector Machines.

4.1 LINEAR SUPPORT VECTOR MACHINES (LSVM)

Given a training set $S = \{(x_i, y_i)\}_{i=1}^{n}$, for n nodes. Support vector learning tries to find a hyperplane, determined by a vector w with minimal norm and an offset vector b, that separates the training data {xi} into two classes denoted by yi={-1,+1}. Let SVM = {w,b} denote the separating hyperplane. To find such a hyperplane, one must solve the following quadraticproblem [11]:

$$\begin{array}{c} min \\ w, \pounds \end{array} \varphi(w, \pounds) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\pounds_i \qquad (1)$$

Subject to

$$y_i(w.x_i + b) \geq 1 - \pounds_i\ \text{and} \quad (2)$$

$$\pounds_i \geq 0, \text{for}\ i\ =\ 1,2,\ldots n$$

Where b determines the offset of the plane from the origin, the set of variables $\{£_i\}_{i=1}^n$ measures the amount of violation of the constraints, and C is a parameter that defines the cost of constraint violation. The vector of minimal norm w that represents the resulting separating hyper plane

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \qquad (3)$$

is expressed by means of a linear combination of the so-called support vectors, i.e., the training sample vectors $\{x_i\}_{i=1}^l$ corresponding to the l non-zero Lagrange multipliers $\{\alpha_i\}_{i=1}^l = 1$, calculated during the optimization process. In practical settings, the number of support vectors is usually quite small compared to the number of training samples (<< n). The decision function for classifying a new point x can be easily written as

$$f(x) = \sin(\sum_{i=1}^{l} \alpha_i y_i x.x_i + b) \qquad (4)$$

and the corresponding decision rule can be expressed as follows: a new test vector xbelongs to class 1 when f (x) > 0 while x belongs to class -1 when f (x) < 0.

4.2 NONLINEAR SUPPORT VECTOR MACHINES AND KERNELS

The nonlinear decision surfaces is necessary because the real-life classification problems are hard to be solved by a linear classifier[12]. When the decision function is not a linear function, the nodes will be mapped from the input space into a high dimensional feature space by a nonlinear transformation. This nonlinear transformation is performed in an implicit way through so-called kernel functions. Table 1 shows three typical kernel functions that can be used for this task.

Table 1: Summary of inner-product kernels[13]

Kernel function	Inner Product Kernel $k(x, x_i), i = 1,2,..N$
Polynomial kernel	$k(x, x_i) = (x^T x_i + 1)^d$
Gaussian (Radial-basis) kernel	$k(x, x_i) = \exp(-\|x - x_i\|^2 / 2\alpha^2)$
Multi-layer perception (sigmoid)	$k(x, x_i) = \tan(\beta_0 x^T x_i + \beta_1),$ β_0, β_1 are decided by the user

Selection of the appropriate kernel function is an important research branch in the area of SVMs. In practical applications, free parameters related to the selected kernel function may generalization performance of SVM. Although there are some theoretical results about the lower and upper bound of these parameters[15], these methods are too complex to be put into practice. Hence, for a given kernel function, selection of relative parameters demands excessive attentions in order to achieve desired level of generalization. In Gaussian (RBF) case, a radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, and the SV algorithm automatically determines centers, weights and threshold such that minimizing an upper bound on the expected test error[14]. Similarly, the Gaussian kernel SVM shows good features and strong learning capability. Our goal is to enable SVM in an efficient way so that: a) we can get good classification results with minimum data redundancy. b) Our algorithms can be used easily in the context of WSN. So we propose the RBF case in our algorithm due to its good features where α is a free parameter.

Generally, K-SVM classification algorithm contains two main phases. First phase is training, which accepts as input the classified nodes groups generated by k-means algorithm and produces a set of weights which will be used during the next phase. The second phase, classification, uses the weights created during training to assign a discriminate or score to each node. Based on this score each node is classified in or out the class.

Figure 4 is the proposed algorithm in its two phases: training and test phases. The figure illustrates the training phase of algorithm k-SVM, where given a set of training samples $\{x_i\}_{i=1}^l$, where$x \in S$, the clustering centers set $\{v_i\}_{i=1}^K$, where K is the number of samples is yielded by the k-means. Then using these centers as the parameters of cluster head selection. In the last step, the RBF classification function for the samples with the transformed features is established using the SVM criterion. The predicted class label y of x node is computed using the trained classification function f(x)yielded in the last step of the training phase.
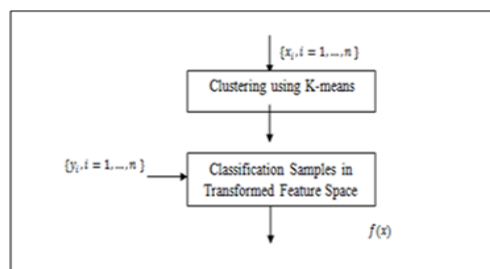

Figure 4: Flowchart of K-SVM Algorithm

5. EXPERIMENTAL ANALYSIS

We have stimulated the clustering process using a library for Support Vector Machines "LIBSVM" included by MATLAB system[16].

From [10], authors show the optimum number of clusters in WSN is around 3–5 for the 100-node network. When there are only one or two clusters, the non-cluster head nodes often have to transmit data very far to reach the cluster head node, draining their energy. Also, when there are more than five clusters, there is not as much local data aggregation being performed. So we suggest for our experiments, we set number of clusters to be five clusters.

In the sequel, method-SVM is developed using MATLAB. As simulation experiment setup, 100*100 sensor area with 100 nodes employed is generated randomly and is used as the training and testing data, with different percentages. Throughout excrement, the following hypothesis are assumed to be hold:

 (1) The battery power of all nodes is randomly equipped with a battery verified from 2 to 5 Joule.

 (2) The communication among the cluster heads CH can cover all nodes in the WSN, that is, each node can be accessed from any other node via the cluster heads.

 (3) The election of CH is not periodic, instead it will be chosen once at the begging.

 (4) During the execution of the algorithm, the network topology structure does not change.

 (5) The base station is located at position (50,50).

From algorithm 2 in figure 5, with the RBF case, because the spread parameter γ where $\gamma = \frac{1}{2\alpha^2}$ and $\gamma > 0$ is closely associated with generalization performance of SVM in LIBSVM, how to choose an appropriate γ is worth pursuing. In our excrements, in training phases, selection of the optimal γ parameter with the RBF kernel for classification problems and adaptive determination of an appropriate for classification in WSNs are studied and some valuable conclusions are obtained as shown in table 2.

The obtained data used for the cauterization consisted of 100 sessions, of which 80 times were used for training and the remaining 20 times for the evaluation of K-SVM performance. In each session, we use 70 nodes as training samples and 30 nodes as testing samples.

Algorithm 2: K-SVM Classifier

Input: $X = \{x_1, x_2, \ldots\ldots, x_n\}$ (set of nodes clustered by K-means)

k (number of clusters)

Output :

$predicted\ class\ label\ y\ for\ x\ node$

- *Transform x nodes to the format of an SVM package*

- *Conduct simple scaling on the data.*

- *Consider the RBF kernel:* $k\,(x\,,x_i\,) = exp\,(-||x - x_i||^2 / 2\alpha^2).$

- *Use cross-validation to find the best parameter γ.*

- *Use the best parameter γ to train the whole training set.*

- *Test.*

Figure 5.K-SVM Algorithm

Using the method referred to as an incremental learning with the SVM[17], we employed measure of K-SVM classifier, namely: accuracy. It is one of the standard metrics commonly used in the context of multiclass problems. Each metric with respect of test examples is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} * 100 \qquad (5)$$

Where TP stands for true positive, TN true negative, FP for false positive and FN for false negative classification in WSN. For example, when the random given clusters are concerned, TP represents the obtained clusters as good, TN represents the obtained clusters in a false way. Otherwise, FP represents misclassified clustering and FN represents non-clusters being misclassified as non-clusters.

Table 2: Testing Results

Session Number	LSVM	RBF Kernel	
	Accuracy	γ	Accuracy
1	29%	9.5367e-07	26%
2	29%	1.9073e-06	26%
3	29%	3.8147e-06	26%
4	29%	7.6294e-06	49%
5	33%	1.5259e-05	80%
6	50%	3.0518e-05	80%
7	61%	6.1035e-05	88%
8	88%	*0.00012207	94%
9	90%	0.00024414	91%
10	91%	0.00048828	94%
11	90%	0.00097656	94%

From Table 2, it can be seen that when γ has a small value (e.g., γ = 9.5367e-07), the accuracy rate is small (26%). Then, the accuracy rate increases with an increment of γ and remains small value within [26%,49%] during a certain range of γ (i.e.,[9.5367e-07, 7.6294e-06]). After $\gamma >$ 1.5259e-05, the accuracy rate varies mildly. When γ = 0.00097656, accuracy reaches 94%. By the presented approach, the optimal γ is 0.00012207. The number of support vectors reaches the least and the accuracy rate is nearly the largest. A similar trend LSVM affected by the number of support vectors can also be seen from Table 2. Accuracy rate of them can hardly up to 90%. Figure 6 shows the accuracy of the clustering process with the γ parameter.
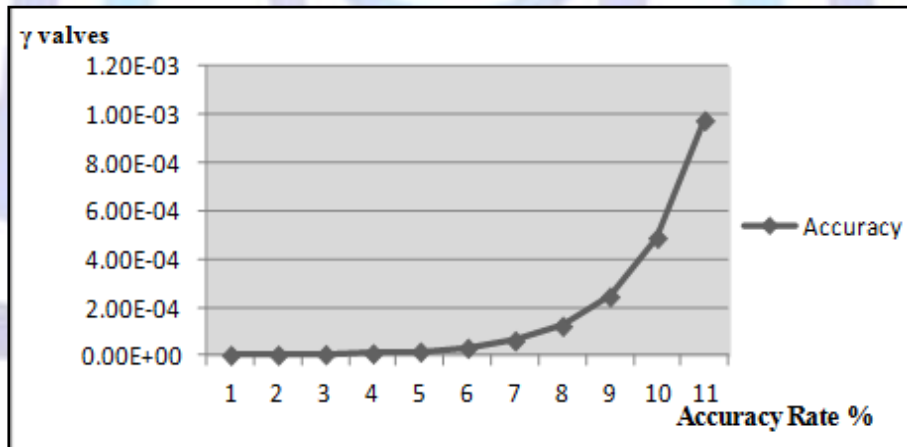

Figure 6: The Accuracy of the Clustering Process in WSN based on K-SVM with RBF Kernel.

Figures 7 and 8 show the exemplary output of the cauterization process with the LSVM and K-SVM with RBF kernel, respectively, for a sensor network of 100 sensors that are distributed uniformly in a square of 100 square units and read nodes refer to cluster heads. It was observed that the clustering process in our proposal gives better results.
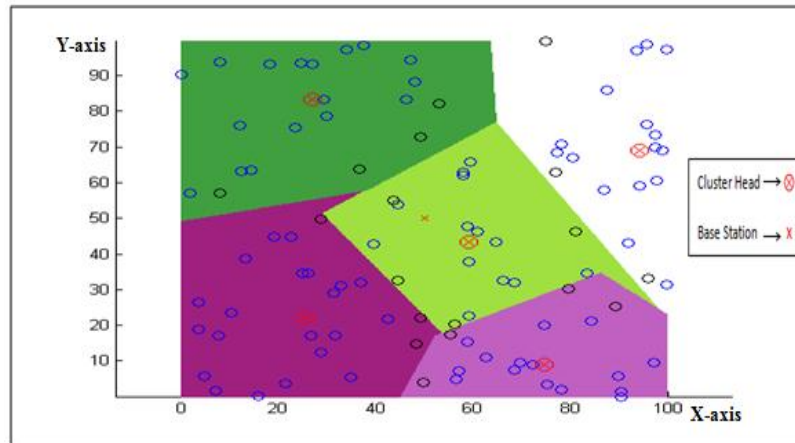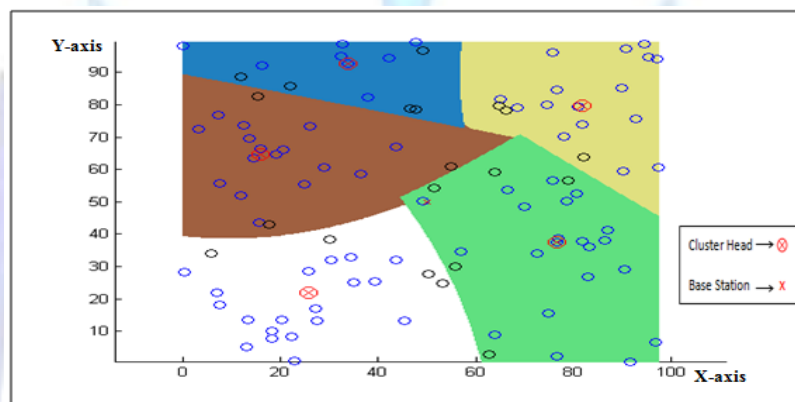
Figure 7: The Clustering Process with LSVM

Figure 8: The clustering Process with the RBF Kernel and $\gamma = 0.00012207$

6. CONCLUSION

In this paper , we use the advantages of Support Vector Machine SVM to strengthen K-MEANS clustering algorithm which gave us a new method called K-SVM .Experimental results show that the proposed method K-SVM using RBF as kernel function gave us more accurate dissection boundary for each classes in the network that reduce data redundancy and extend network lifetime when compared with typical K- means algorithm.

REFERENCES

[1] "A Survey on PowerControl Issues in Wireless Sensor Networks," IEEE Com-munication Surveys, Vol. 9, No. 4, 2007, pp. 86-107.

[2] I. F. Akyildiz, W. Su, Y. SankarasCayirci, "Wireless Sensor Networks: A Survey," Com-puter Networks, Vol. 38, No. 4, 2002, pp. 393-422. doi:10.1016/S1389-1286(01)00302-4.

[3] A. Giridhar and P. R. Kumar, "Maximizing the Functional Lifetime of Sensor Networks," Fourth InternationalSymposium on Information Processing in SensorNetworks, Los Angeles, 15 April 2005, pp. 5-12.

[4] S. P. Lloyd, "Least-squares quantization in PCM," IEEE Trans. Inf. Theory, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[5] S. Basagni, "Distributed Clustering Algorithm for Ad-Hoc Networks," Proceedings of International Symposium

on Parallel Architectures, Algorithms, and Networks (I-SPAN), Perth, 1999, pp. 310-315.

[6] S. Banerjee and S. Khuller, "A Clustering Scheme for Hierarchical Control in Multi-Hop Wireless Networks," Proceedings of IEEE INFOCOM, Anchorage, 2001, pp. 1028-1037.

[7] C.R. Lin, M. Gerla, Adaptive clustering for mobile wireless networks, IEEE Journal on Selected Areas in Communications 15 (7) (1997) 1265– 1275.

[8] ZdenekHanzalek, and PetrJurcik, ''Energy Efficient Scheduling for Cluster-Tree Wireless Sensor Networks With Time-Bounded Data Flows: Application to IEEE 802.15.4/ZigBee'', IEEE TRANSACTIONS on industrial informatics, vol. 6, no. 3, august 2010.

[9]J. Martyna, "An Application of LS-SVM Method for Clustering in Wireless Sensor Networks," in New Challenges in Applied Intelligence Technologies, N. T. Nguyen and R. Katarzyniak, Eds. Springer Berlin Heidelberg, 2008, pp. 383–392.

[10]W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," IEEE Transactions on wireless communications, vol. 1, no. 4, pp. 660–670, 2002.

[11]K. Bennett and C. Campbell, "Support vector machines: hype or hallelujah," SIGKDD Explorations, vol. 2,2, pp. 1–13, 2000.

[12]F. Tay and L.J. Cao, Modified support vector machines in financial time series forecasting, Neurocomputing, Sept., (2001).

[13] S. Haykin, Neural Networks, Prentice Hall Inc. (1999).

[14]B. Scholkopf, K.-K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2758–2765, 1997.

[15] V. Vapnik, The Nature of Statistical Learning Theory, Wiley, New York, USA, 1995.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, p. 27, 2011.

[17] Ruping, S.: Incremental Learning with Support Vector Machines. In: Proc. IEEE Int. Conf. on Data Mining, San Jose, CA, USA, November 2001, pp. 641–642 (2001)

[18] P. Sasikumar and S. Khara, "K-Means Clustering in Wireless Sensor Networks," in 2012 Fourth International Conference on Computational Intelligence and Communication Networks (CICN), 2012, pp. 140–144.

[19] M. N. Joshi, "Parallel k-means algorithm on distributed memory multiprocessors," Computer, vol. 9, 2003.

Author' biography with Photo

Dr Manal received her PhD at computers and systems engineering, Faculty of engineering, Ain Shams University, Cairo, Egypt, 2002. She has experience in industrial computer networks and her research interest includes computer networks, performance evaluation, WSN, network management, and BD management. Currently she is assistant professor, Faculty of Computing and Information Technology FCIT female section, King Abdulaziz University, KAU, Saudi Arabia.

Ahlam S. Althobaiti is Master's student in computer Faculty of Computing and Information Technology FCIT King Abdulaziz University, Saudi Arabia. Her research interests lie in the general field of the computer networks with special focuses on wireless applications and problems in enterprises security.Currently she is teaching assistant,Computer Science Department,College of Computers and Information Sciences,Taif University,Taif, Saudi Arabia.

Afnan B. Bawazeer received her Bachelorat computers science , Faculty of Computer and information system, Taif University, Saudi Arabia, 2010. She has experience in computer networks and computer vision. Her research interest includes computer networks and experimental algorithms. Currently she is Master student, Faculty of Computing and Information Technology, KingAbdulazizUniversity, KAU, Saudi Arabia.

Nouf H. Al-Harbe received her BS degree at Computer Science, Faculty of Computer Science and Engineering, Taibah University, Saudi Arabia. After graduation she worked in Taibah University as Teaching Assistant. Currently, she is MS student in the Computer Science Department, King Abdulaziz University, KAU, Saudi Arabia.