Data Mining of Bayesian Networks to Select Fusion Nodes from Wireless Sensor Networks

Yee Ming Chen[1] Chi-Shun Hsueh[2] Chu-Kai Wang [3]

[1,3] **Department of Industrial Engineering and Management, Yuan Ze University**
135 Yuan-Tung Rd., Chung-Li, Tao-Yuan, Taiwan, R.O.C.
[2] **Information and Communication Research Division,**
National Chung Shan Institute and Technology , Tao-Yuan, Taiwan, R.O.C.

Abstract

In the wireless sensor networks(WSN) management, under different regions and different time, spectral data is very big, resulting in the sensor nodes management would be difficult, this paper based on Bayesian network of data mining to select fusion node in the wireless sensor networks, which focused on balancing energy consumption. Under the architecture of Bayesian network, in which can integrate the quantitative and qualitative knowledge into a comprehensive probabilistic knowledge representation and inference WSN environment. It discusses how these techniques can prepare mining the wireless sensor networks data inside the network (in-network) before data fusion further processing as big data.

Keywords: *Data mining, Bayesian Networks, Data fusion*

1. Introduction

Wireless sensor networks (WSN) consist of a collection of a large number of small, distributed over a large area, and low powered sensor nodes capable of communicating with each other via an ad-hoc wireless network. Each WSN consists of three primary components: sensor nodes, fusion node, and a remotely base station. The sensor nodes are responsible for collecting the locally available sensor data. The sensor nodes are small and inexpensive. Since most nodes are traditionally battery powered, power consumption is an important consideration when setting up a WSN. Once the data has been collected from the sensor nodes, they then transmit that information to a fusion node [1]. The brain of a WSN is usually a decision-making algorithm that is capable of correctly mapping a set of newly collected observations from the sensors to one or more predefined fusion node [2]. Due to their limited power and short communication range, the information that the nodes send to the remotely base station is usually put through a data fusion technique before being sent to the base station. This allows for the data to be more accurate and also reduces overhead in the network[2].

The need for extracting knowledge from the sensor data, collected from WSN, has also become an important issue in real-time decision algorithms. In addition, the rapid change of the monitored data requires the implementation of data mining algorithms in order to get a reasonable time response or prediction. An imbalanced energy problem may have another implication in WSN where it could be a symptom of producing traffic hot-spot in WSN. The energy consumption in the sensors may become imbalanced too, which leads to drain out for some local nodes. Some data mining solution has been selected to better fusion nodes and traffics although it is aimed at the energy level. The data mining algorithms could be generally classified into either a centralized or a distributed data processing [3]. As a distributed data processing model for the probabilistic relationships among a set of variables, the Bayesian network (BN) has become a popular representation for encoding uncertain expert knowledge in data mining domains over the last decade [4]. This paper use data mining approach, Bayesian Networks, to select a fusion node.

The outline of this paper is as follows. Distributed sensors data mining processing approaches are presented in Section 2. Developed fusion node selection using Bayesian networks are stated in Section 3. Simulation results on data fusion discover fusion node are presented in Section 4. Finally, conclusions are made in Section 5.

2. Distributed Sensors Data Mining Processing Approaches

In the distributed sensors data processing approach, each node uses its limited computing resources to perform the mining process. The process of acquiring the data is another factor impacting the data quality, because the transmission of the signal is affected by all kinds of uncertain factors. Thus, uncertainty is an inherent property of the WSN data. The Bayesian method integrates a prior knowledge about the targets under study and the information provided by new data set, followed by encoding the multi-knowledge into conditional probability network model. Thus, Bayesian network in conjunction with Bayesian statistical techniques facilitates the combination of domain knowledge with the relevant data.

The syntax of Bayesian network is clearer, which can reason in dual direction and can be constructed and debugged rapidly. The disadvantage of Bayesian network is that the computation complexity is high. Bayesian Network Syntax illustrated as the following [5]:

BN= (Structure, CPT)

(1) Structure contain nodes and arcs

Nodes: random variable.
(a) Nodes can be continuous or discrete.
(b) Nodes can have two states or more.
(c) Nodes can be deterministic or nondeterministic.

Arcs: relationships between nodes.
(a) Arcs represent causal relationships of nodes.
(b) Arc between x and y represents that x has direct causal influence only.

(2) CPT: Condition Probability Table
(a) Each node has condition probability which is stored in a table (CPT).
(b) Value in table is P(Xi|parents(Xi)), parents(Xi) is the set of parent nodes of Xi.
(c) Root node is particular, as it has no parent node and has only prior probability: parents $(X_i) = \Phi$, so $P(X_i \mid parents(X_i)) = P(X_i)$.

Suppose we have two variables E and H. If H has been realized (i.e. we know exact value of H), we might want to know what is probability of the event E. The situation where we are dealing with probability of one event, given that another event has occurred, is called conditional probability. Mathematically conditional probability is defined:

$$P(E|\, H_k) = \frac{P(E \cap H_k)}{P(H_k)} \; . \qquad (1)$$

If $E_1, E_2, \ldots, E_n$ are n mutually exclusive events such that , $\bigcup_{i=1}^{n} E_i = \Omega$, E_i's are said to be exhaustive. Two variables are said to be disjoint if they have no elements in common. If variables are disjoint and exhaustive Equation (2) holds:

$$E = \bigcup_i (E \cap H_i)$$
$$(E \cap H_i) \cap (E \cap H_j) = \emptyset, i \neq j \qquad (2)$$

The theory of the Bayesian networks assumes that events are disjoint and exhaustive. If they are not, the results are not consistent. When events are disjoint and exhaustive, the probability of E can be calculated via conditional probabilities:

$$P(E) = \sum_{i=1}^{n} P(E \cap H_i) = \sum_{i=1}^{n} P(E \mid H_i) \cdot P(H_i) \qquad (3)$$

Using Equation (1) we can express the sum of the intersection of E and H as follows.

$$P(E \cap H_k) = P(E \,|H_k) \cdot P(H_k) = P(H_k|E) \cdot P(E) \qquad (4)$$

Now it is possible to place Equation.(3) to Equation.(4) and obtain Equation (5).

$$P(H_k|E) = \frac{P(E|H_k) \cdot P(H_k)}{P(E)} = \frac{P(E|H_k) \cdot P(H_k)}{\sum_i^n P(E|\, H_i) \cdot P(H_i)} \qquad (5)$$

Equation (5) is widely known as Bayes' formula. Bayes' formula is the foundation of Bayesian networks that are actually nothing but a network structure where observations are handled using Bayes' theorem. In the Equation (5), H_k means all hypotheses (subscript *k* refers to the fact that there are several hypotheses). Probabilities are a priori values from experts. A term $P(H_k)$ is called a priori probability. The denominator can be considered as a normalizing factor which normalizes the probability between zero and one.

3. Fusion nodes selection in WSN using BN

For WSN, these sensor nodes not only detect the target, but they also collect the data process and transmit it to the outside world for further processing. These sensor nodes require careful resource management as they are tightly constrained in terms of power, transmission power, processing capacity and storage capability. The data being sensed by each sensor node must ultimately be transferred to a remotely base station. The communication is expensive in terms of energy usage between sensor nodes and base station. A fusion node is selected in each WSN to communicate with the remotely base station on behalf of other sensor nodes in the efficient manner [6]. Once a sensor node is selected a fusion node, its overall energy consumption increases significantly as it has to communicate with all other sensor nodes within coverage of region of interest (ROI) as well with the remotely base station. Therefore the process continues periodically and in each round different fusion node selected to balance the energy consumption throughout the WSN. Fusion node selection is an important procedural step due to the exact location of sensor node. If the selected fusion node is located closer to most of the sensor nodes within ROI, the cost of communication between fusion node and other nodes will be minimal. On the other hand if fusion node is located far away from majority of the sensor nodes then the transmission cost from each sensor node to fusion node will be higher and the other sensor nodes will consume more energy as stronger signal will be required to communicate [7]. The probability of each sensor node becoming a fusion node based on its probabilistic distance from remaining sensor nodes is computed and the one with the highest probability is selected as the fusion node. Since the exact location which is required for fusion node selection is not known, therefore each node calculates the distance probabilistically. The distance depends upon the information a node receives from all other node in the WSN. This information comprises of the signal strength and energy. Bayesian network is used to calculate the probabilistic distance between each pair of nodes. The probabilistic distance together with residual energy of each node enables the Bayesian network to find the most probable candidate to become a fusion node. The whole process is repeated to obtain new fusion node after each round and the residual energy is adjusted accordingly after every round.

Since a WSN possibly consist of from tens to hundreds of sensor nodes, size of WSN depends upon type of application where the network is employed. In this paper as we are focusing only on the fusion node selection process, a small BAN is used that consist of five nodes scattered randomly in an area. Figure 1 shows the scenario used as the example. The figure shows a scenario where five sensor nodes are located in a BAN and are scattered randomly. From the Figure 1 it is evident that sensor node number 5 is located far from the rest of the four other nodes. Sensor node 1, 2 and 3 are closed to each other and node 4 not far from node 2 and 3. In our research since location of sensing nodes are not known to the system, therefore we have employed Bayesian Network to find the probabilistic distance among the nodes in the WSN.

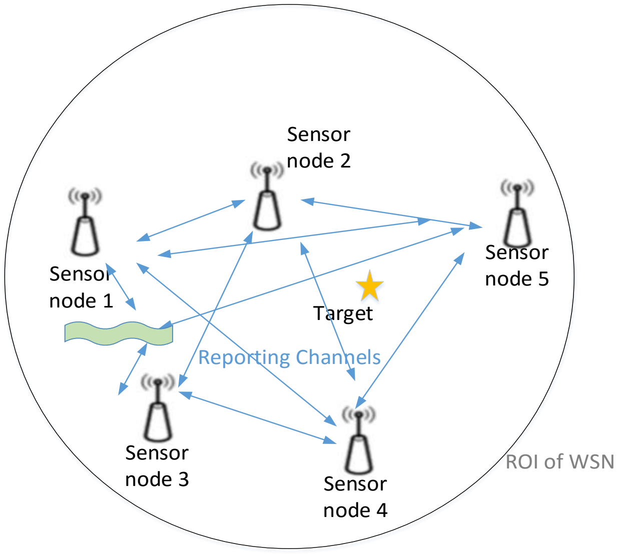

Figure 1 The Sensor nodes scattered randomly in a ROI of WSN

4. Data mining discovery fusion node

In this section, we formulate Bayesian Network model to discover the fusion node. At the start of each round, each sensor node can potentially become a fusion node, the selection depends upon the location of the node with respect to rest of the sensor nodes in the ROI of WSN and the residual energy level. Therefore at the start of each round each node will calculate its distance probabilistically from all other sensor node in the WSN, the distance depends upon signal strength and power level. As every sensor node in the WSN becomes a fusion node therefore Bayesian Network must be built for every node considering it to be a candidate for fusion node. An example of such network for node N1 is shown in Figure 2 below;

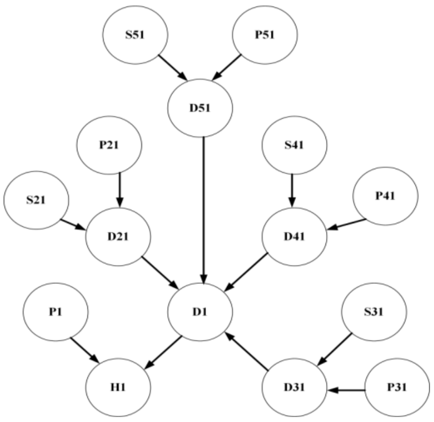

Figure 2 Bayesian Network for Sensor node 1.

Figure 3 below shows that how sensor node 1 sees the other sensor nodes in the WSN from its own perspective. H1 is the probability of sensor node N1 to become a fusion node, this probability is calculated by running the simulation using Netica$^{@}$. D1 is the probabilistic distance that shows how close node N1 is from all other nodes within the WSN. D21 is the distance from sensor node N2 to N1, D31 is the distance from sensor node N3 to node N1 and so on. S21 is the signal strength received by sensor node N1. P1 is the power level of sensor node N1 and so on. Probabilistic calculation of D21 is based on the signal strength S21 and power level P2. The same model is repeated for every node in the WSN. Individual nodes will receive probabilistic distance from every other node and then calculates the overall aggregate distance from itself to all other sensor node in the WSN.

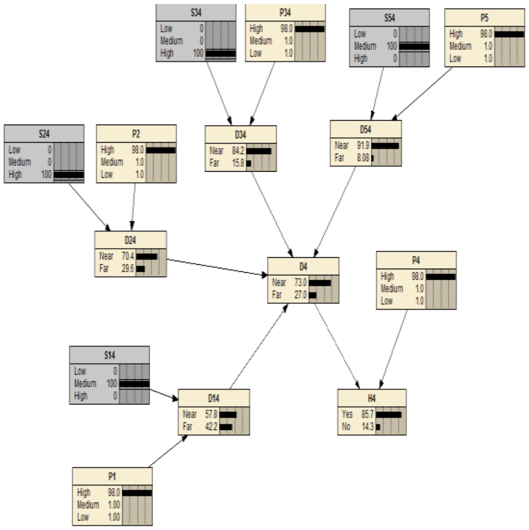

Figure 3 Simulation result from Netica showing probability of sensor node N4 which was selected as fusion node H4 in the first simulation run

Figure 3 above shows the result of simulation for sensor node N4, it can be seen from above that in the present scenario with the available signal strength and power level, H4 has 85.7 % probability of becoming a fusion node for the present round. In the same way each node will calculate its own probability of becoming a fusion node, node showing highest probability will become fusion node for the current round. In our simulation following assumptions have been made:

WSN coverages are already formed before the fusion node selection process and the WSN's ROI could be of different size. The assumptions list as the following:

(1) The remotely base station is located far from the WSN coverage;

(2) All sensing nodes are immobile;

(3) All sensing nodes are homogenous, and are energy constrained;

(4) Sensor nodes have no location information;

(5) After every round each other nodes will consume 5% energy;

(6) Every fusion node after each round will consume 8% of energy.

Sensor node N5 in the present scenario is located comparatively far from the rest of the sensor nodes therefore it consume more energy, 6% of energy will be consumed by sensor node N5 after each round.

Figure 4 below shows the result of the fifth simulation for sensor node N1, it can be seen from above that in the present scenario with the available signal strength and power level, node 1 has 72.8% probability of becoming a fusion node (H1) for the fifth round. In the same way each node will calculate its own probability of becoming a fusion node, node showing highest probability will become fusion node head for the current round. At the end of each round the node showing the highest probability among the five sensor nodes will become a fusion node for that particular round.

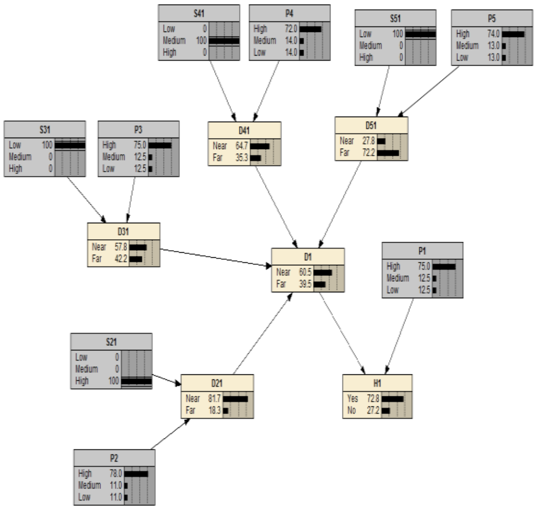

Figure 4 Simulation result from Netica showing probability of sensor node N1 which was selected as fusion node H1 in the fifth simulation run

Table 1 below shows the simulation results, at the start of round 1, each sensor node carries 98% energy, from P1 to P5. The probability of each node is listed from H1 to H5. At the end of each round, new power level value is calculated for next round, reduced power level is recorded and new fusion node is selected accordingly. As can be seen from the entry in the table that for round 1, node H4 has the highest probability among the five nodes and hence selected as fusion node for round 1.

Table 1 Simulation results from Netica using Bayesian Network Approach

Rounds	P1	P2	P3	P4	P5	H1	H2	H3	H4	H5	Candidate Fusion node
1	0.98	0.98	0.98	0.98	0.98	0.82	0.82	0.84	0.86	0.77	H4
2	0.93	0.93	0.93	0.90	0.92	0.79	0.80	0.82	0.82	0.75	H3 or H4
3	0.88	0.88	0.88	0.82	0.86	0.77	0.78	0.80	0.79	0.73	H3
4	0.83	0.83	0.80	0.77	0.80	0.74	0.75	0.72	0.73	0.71	H2
5	0.75	0.78	0.75	0.72	0.74	0.73	0.70	0.69	0.69	0.70	H1

5. Conclusions

In this survey paper, we address the problems of fusion node selection in wireless sensor network. Due to the wireless sensor networks is large, and it corresponds to different time and place, the record of the spectral information of data is very huge, in such big data condition. Uncertainty is an inherent property of WSN Sensed data. Under the architecture of Bayesian network, which can integrate the quantitative and qualitative knowledge into a comprehensive probabilistic knowledge representation and inference environment, this paper presents a BN model for data mining from sensing data.

Acknowledgments

This research work was sponsored by the Ministry of Science and Technology, R.O.C., under project number MOST104-2221-E-155-019.

References

[1] Y. Gai, B. Krishnamachari and R. Jain, “Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation”, New Frontiers in Dynamic Spectrum, IEEE Symposium on IEEE, 2010, pp. 1-9.

[2] Davood Izadi, H. Jemal, Sara Ghanavati Abawajy, and Herawan Tutut. "A Data Fusion Method inWireless Sensor Networks." Sensors, vol.15, no. 2 (2015), 2964-2979.

[3] Yee Ming Chen and Wen-Yuan Wu “Cooperative Electronic Attack for Groups of Unmanned Air Vehicles based on Multi-agent Simulation and Evaluation,” International Journal of Computer Science Issues, vol. 9 (2), 2012, pp. 11 – 16.

[4] Z. Han, R. Zheng and H. V. Poor, “Repeated auctions with Bayesian nonparametric learning for spectrum access in cognitive radio networks”, Wireless Communications, IEEE Transactions on, vol. 10, no. 3, 2011, pp. 890-900.

[5] X. Xing, T. Jing and Y. Huo, “Channel quality prediction based on Bayesian inference in cognitive radio networks”, INFOCOM, Proceedings IEEE, 2013, pp. 1465-1473.

[6] Bill C.P. Lau, Eden W.M. Ma, Tommy W.S.Chow, “Probabilistic fault detector for Wireless Sensor Network”, Expert System With Applications , Volume 41, Issue 8,15 , 2014, pp 3703-3711.

[7] X. Xing, T. Jing and W. Cheng, “Spectrum prediction in cognitive radio networks”, Wireless Communications, IEEE, vol. 20, no. 2, 2013, pp. 90-96.