ĐỀ XUẤT GIẢI PHÁP TIỀN XỬ LÝ ĐỂ TỔNG HỢP DỮ LIỆU NHIỀU CẢM BIẾN TRONG MẠNG CẢM BIẾN KHÔNG DÂY

Dương Viết Huy[1], Nguyễn Đình Việt[2]

[1] Vụ Khoa học, Công nghệ và Môi trường - Bộ Văn hóa, Thể thao và Du lịch
[2] Đại học Công nghệ, Đại học Quốc gia Hà Nội

huy.duongviet@gmail.com, vietnd@vnu.edu.vn

TÓM TẮT *- Giải pháp chia mạng cảm biến không dây (WSNs - wireless sensor networks) thành nhiều cụm (cluster), mỗi cụm có nhiều nút cảm biến (multi-sensor) để tổng hợp dữ liệu tại các nút trung gian trên đường truyền từ nút cảm biến về mục tiêu đến trạm đích (BS - base station) đang được nhiều nhóm nghiên cứu. Tổng hợp dữ liệu nhằm hạn chế các gói tin dư thừa do các nút cảm biến trong cụm cùng cảm nhận về một đối tượng nên thường có cùng thông tin và cùng truyền dữ liệu này đến BS gây tổn hao năng lượng vô ích đồng thời tăng nguy cơ nghẽn đường truyền đến BS. Tại mỗi cụm có một nút cụm trưởng (CH- cluster head) chịu trách nhiệm tổng hợp dữ liệu từ các nút trong cụm đó gửi đến BS. Một trong những yếu tố quyết định hiệu quả của việc tổng hợp đó là chất lượng dữ liệu đầu vào mà CH nhận được từ các nút trong cụm gửi về. Do nút cảm biến thu phát tín hiệu bằng sóng điện từ nên sẽ có rất nhiều yếu tố ảnh hưởng đến việc đo lường về mục tiêu như nhiễu, mất dữ liệu... Nếu CH sử dụng ngay kết quả đo này làm dữ kiện đầu vào để tổng hợp dữ liệu thì có thể không phản ánh đúng sự kiện diễn ra ở mục tiêu. Bài báo này đề xuất giải pháp tiền xử lý DP-DF nhằm loại bỏ dữ liệu thô, giữ lại dữ liệu có nhiều giá trị về tri thức tham gia tổng hợp dữ liệu.*

Từ khóa *- Tổng hợp dữ liệu, tiền xử lý, multi-sensor, data fusion, DP-DF, WSNs.*

I. GIỚI THIỆU

Hiện nay, hệ thống giám sát bằng mạng cảm biến ngày càng phát triển về quy mô (số nút cảm biến, phạm vi giám sát) và chất lượng (số tham số giám sát, độ mịn của mức đo,…). Thông thường, các nút cảm biến không dây được "nuôi" bởi nguồn pin hữu hạn, do vậy khi nghiên cứu về WSNs thì vấn đề tiết kiệm năng lượng của nút và của mạng luôn được đặt ra. Một trong những nhóm giải pháp được nhiều nhóm nghiên cứu đó là mạng có phân cụm (*cluster-based network*). Giải pháp phân cụm, điển hình là công trình [1] với mục tiêu chia nhỏ mạng cảm biến thành các mạng cơ sở còn gọi là cụm (*cluster*), giao tiếp trong cụm có thể theo kiểu đơn chặng – *singlehop* hoặc đa chặng - multihop. Nút trưởng cụm (*CH - cluster head*) chịu trách nhiệm tổng hợp dữ liệu (*data fusion* hoặc *data aggregation*, chúng tôi sẽ sử dụng thuật ngữ *data fusion - DF*) đồng thời tham gia quá trình định tuyến. Sau mỗi vòng, mạng phải phân chia lại thành các cụm mới và phải bầu ra CH mới để tiếp tục hoạt động.

Các nghiên cứu [2, 3] đã đề xuất giải pháp tổng hợp dữ liệu nhiều cảm biến tại nút CH dựa vào bảng dữ kiện của thuộc tính ngữ nghĩa. Tại thời điểm DF, dữ liệu cảm nhận của các nút cảm biến trong cụm được hệ thống hóa thành bảng thông tin ngữ nghĩa gồm *ngữ nghĩa của nút cảm biến* (như khoảng cách, năng lượng còn lại,...) và *ngữ nghĩa của dữ liệu cảm nhận* (như độ chính xác, số gói tin cần truyền,…). Từ các kết luận về ngữ nghĩa, CH sẽ lựa chọn nút cảm biến thỏa mãn điều kiện *để chuyển tiếp dữ liệu cảm nhận* của nút cảm biến đó đến BS.

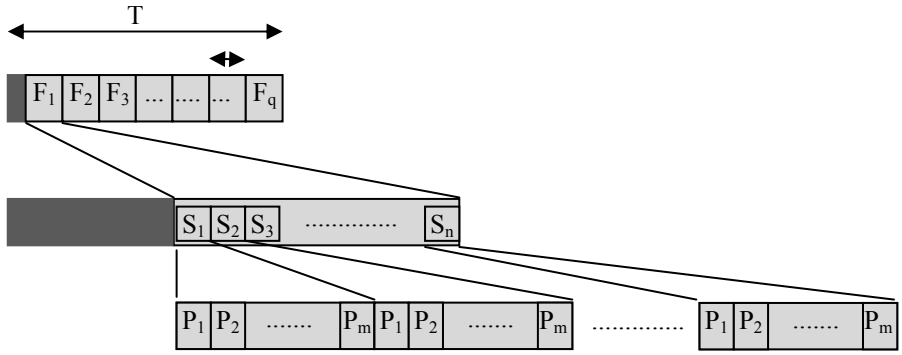
Vì các nút cảm biến thu phát tín hiệu bằng sóng vô tuyến nên chúng luôn tiềm ẩn nhiều tình huống làm giảm chất lượng dữ kiện đầu vào như dữ liệu không chắc chắn, bị thiếu, dữ liệu yếu,... ảnh hưởng đến quá trình tổng hợp và kết quả dữ liệu đầu ra tại nút CH. Do đó, trước lúc DF, dữ liệu cần phải được xử lý. Giai đoạn tiền xử lý trong bài toán tổng hợp dữ liệu nhiều cảm biến được tính từ lúc các nút cảm biến trong cụm cảm nhận mục tiêu và gửi đến CH đến lúc CH đóng dữ liệu thành khối dữ kiện đầu vào để tiến hành tổng hợp trước khi gửi đến BS.

Trong bài báo này, chúng tôi đề xuất phương pháp tiền xử lý dữ liệu với tên gọi *DP-DF (Data Pre-processing for Data Fusion)* bằng việc *áp dụng entropy thông tin và lý thuyết tập thô* nhằm chuẩn hóa dữ liệu đầu vào của các nút cảm biến trong cụm gửi về CH phục vụ tổng hợp dữ liệu nhiều cảm biến tại nút CH. Nội dung bài báo ngoài giới thiệu và kết luận có 2 nội dung chính: Phân tích giai đoạn tiền xử lý dữ liệu phục vụ tổng hợp dữ liệu nhiều cảm biến; đề xuất giải pháp DP-DF và ví dụ minh họa quá trình tiền xử lý đã đề xuất.

II. TIỀN XỬ LÝ DỮ LIỆU CẢM BIẾN

A. Dữ liệu đầu vào tiền xử lý

Giai đoạn tiền xử lý để tổng hợp dữ liệu nhiều cảm biến (trong mạng cảm biến không dây) trong bài báo này được tính từ lúc các nút cảm biến trong cụm cảm nhận mục tiêu và gửi đến CH đến lúc CH đóng dữ liệu thành khối dữ kiện đầu vào để tiến hành tổng hợp dữ liệu trước khi gửi đến BS. Mục đích của giai đoạn tiền xử lý là hạn chế tối đa các dữ liệu thô, ít có giá trị về tri thức tham gia tổng hợp dữ liệu. Chúng tôi chia thời điểm để đóng gói dữ kiện làm đầu vào để DF thành 2 loại: Theo khung tin *(frame)* hoặc theo chu kỳ/vòng *(T)*. Giả sử mỗi *T* có *q frame (F)*, cụm có *n* nút cảm biến *(S)*, mỗi S đo lường *m* tham số *(P - parameter)*, biểu diễn ở *Hình 1*.



Hình 1. Truyền dữ liệu theo khung tin *(frame)* và theo chu kỳ (T)

1. Theo khung tin

Tại CH, sau khung truyền F_1, CH sẽ nhận được bảng dữ liệu *n* hàng, *m* cột như ở *Bảng 1*.

Bảng 1. Dữ liệu CH nhận của khung truyền F_1

F_1-S_1-P_1	F_1-S_1-P_2	.....	F_1-S_1-P_m
F_1-S_2-P_1	F_1-S_2-P_2	.....	F_1-S_2-P_m
.....	.....	.....	.....
F_1-S_n-P_1	F_1-S_n-P_2	.....	F_1-S_n-P_m

Bảng 2. Dữ liệu CH nhận của khung truyền F_k

F_k-S_1-P_1,	F_k-S_1-P_2	.....	F_k-S_1-P_m
F_k-S_2-P_1	F_k-S_2-P_2	.....	F_k-S_2-P_m
.....	.....	.....	.....
F_k-S_n-P_1	F_k-S_n-P_2	.....	F_k-S_n-P_m

Kết thúc F_1, tại CH, tập dữ liệu để xử lý theo tham số P_j $(1 \le j \le m)$ gồm các phần tử ở cột *j* và tập dữ liệu để xử lý các tham số P_j theo nút cảm biến S_i $(1 \le i \le n)$ là các phần tử ở hàng thứ *i*. Tổng quát, sau khung truyền F_k (với $1 \le k \le q$), CH sẽ nhận được bảng dữ liệu *n* hàng, *m* cột chứa dữ liệu đo *m* tham số của *n* nút cảm biến, mỗi khung truyền sẽ có một bảng. Mỗi phần tử trong bảng là giá trị đo tham số P_j của nút cảm biến S_i, được truyền đến CH ở khung truyền F_k trong chu kỳ truyền *T*, bảng dữ liệu tổng quát như ở *Bảng 2*.

Như vậy, với phương pháp xử lý này, sau khi nhận hết dữ liệu truyền của 1 frame, CH sẽ xử lý với dữ liệu của nút cảm biến và tham số tương ứng trước đó, tích lũy kết quả này để sử dụng khi nhận hết 1 frame liền sau đó. Gọi $F_{k'}$ là kết quả đóng gói sau khi CH nhận hết frame F_k khi đó $F_{k'} = Combine\ (F_k, F_{k-1})$ (1)

$F_{k'}$ có thể được xem là một ma trận cỡ $(n \times m)$ là sự kết hợp tích lũy của 2 ma trận cùng cỡ của F_k và F_{k-1}. Các phần tử của $F_{k'}$ có giá trị là: F_k-S_i-P_j (Với $1 \le k \le q$, $1 \le i \le n$, $1 \le j \le m$). (2)

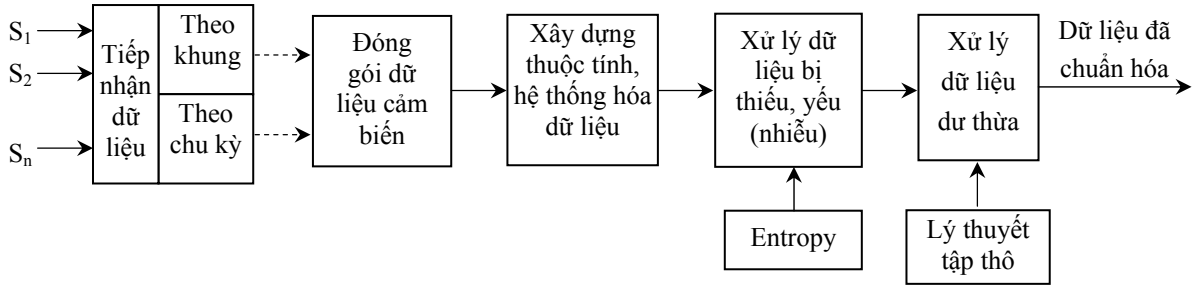
Như vậy, nếu đóng gói theo khung tin thì $F_{k'}$ sẽ là tập dữ liệu đầu vào để CH đóng gói và áp dụng giải pháp tiền xử lý. Kết thúc vòng *(T)* khi $k = q$, lúc này CH nhận hết dữ liệu của *q* khung tin của vòng.

2. Theo chu kỳ/vòng (T)

Tương tự cách diễn giải ở trên, với hình thức này, CH sẽ nhận và lưu đủ dữ liệu của *q* khung tin mới tiến hành đóng gói. Gọi F_{block} là dữ kiện đầu vào để áp dụng giải pháp tiền xử lý, F_{block} bao gồm *q* ma trận cỡ $(n \times m)$.

B. Phân tích tiền xử lý dữ liệu cảm biến

Sau khi CH đóng gói dữ liệu cảm biến theo khung tin hoặc theo chu kỳ, CH sẽ sử dụng dữ kiện này làm đầu vào để áp dụng giải pháp tiền xử lý. Tương tự kỹ thuật tiền xử lý trong khai phá dữ liệu data mining [4], giai đoạn tiền xử lý tại nút CH trong bài báo này gồm các công đoạn và thứ tự xử lý như ở *Hình 2*:



Hình 2. Quá trình tiền xử lý dữ liệu cảm biến tại nút CH của giải pháp DP-DF

- *Xây dựng thuộc tính (attribute/feature construction):* Là các thuộc tính ngữ nghĩa của nút cảm biến và ngữ nghĩa của dữ liệu cảm nhận [2, 3]. Thuộc tính là các cột của bảng dữ liệu cảm biến.

- *Hệ thống hóa dữ liệu:* là quá trình nhận diện đặc điểm chung của dữ liệu cảm biến và sự hiện diện của dữ liệu nhiễu, dữ liệu thiếu hoặc các phần tử kì dị *(outliers)* khi nút cảm biến đo lường; định lượng hóa thành giá trị để đưa vào bảng dữ liệu gồm *n* hàng, *m* cột tương ứng với *n* nút cảm biến của mạng và *m* thuộc tính của mỗi nút cảm biến.

- *Xử lý dữ liệu bị thiếu (missing data):* Khi CH không nhận đủ dữ liệu từ một hoặc nhiều nút trong nhóm để làm dữ kiện cho quá trình DF. Dữ liệu bị thiếu có thể là dữ liệu đo của tất cả các tham số đo về mục tiêu hoặc của một vài tham số đó thành phần của mục tiêu. Do đó, xử lý dữ liệu bị thiếu là bước quan trọng trong giai đoạn tiền xử lý.

- *Xử lý dữ liệu bị nhiễu (noisy data):* Khi nút cảm biến cảm nhận về mục tiêu, tín hiệu có thể bị nhiễu dẫn đến tính chân lý của dữ liệu truyền đi không được bảo toàn. Tiền xử lý tại CH có thể xác định lại sự đúng đắn của dữ liệu cảm nhận bằng cách loại bỏ thông tin nhiễu, giữ lại thông tin hữu ích, ít bị nhiễu để tiến hành DF.

- *Xử lý dữ liệu dư thừa (redundancy):* Đây là một vấn đề rất quan trọng trong bài toán DF. Khi các nút cảm biến cùng cảm nhận về một đối tượng và cùng truyền một loại thông tin đó trực tiếp đến BS hoặc qua nút cảm biến trung gian (là CH nếu mạng có phân cụm) để truyền đến BS thì việc loại bỏ các dữ liệu dư thừa này là điều rất cần thiết. Nghiên cứu [2] là một trong những đề xuất giải pháp ứng dụng lý thuyết tập thô để xử lý dữ liệu dư thừa.

III.GIẢI PHÁP DP-DF

A. Xử lý dữ liệu thiếu, nhiễu

Sau khi kết thúc quá trình đóng gói dữ liệu của n nút cảm biến trong cụm gửi về CH, xây dựng thuộc tính ngữ nghĩa, hệ thống hóa dữ liệu cảm biến, các giá trị ngữ nghĩa được định lượng bằng các giá trị đo và đưa vào bảng dữ liệu, bảng này được xem là một hệ thống thông tin [5] của cụm (có *n* nút cảm biến) ký hiệu *IS* là một bảng dữ liệu gồm *n* hàng, *m* cột - mỗi cột là một thuộc tính, *IS* được biểu diễn bởi 4 yếu tố [5]: IS = <*U, Q, V, f*> (3)

Trong đó, *U* là tập hữu hạn *n* nút cảm biến; *Q* là tập hữu hạn các thuộc tính; *V* là tập giá trị của tập thuộc tính; *f* là giá trị một thuộc tính của một nút cảm biến tương ứng. Hệ thống thông tin *IS* tổng quát tại thời điểm bắt đầu tiền xử lý ở *Bảng 3*. Gọi $f(S_i, A_j)$ là các giá trị *f* của nút cảm biến S_i tại thuộc tính A_j $(1 \le i \le n,\ 1 \le j \le m)$, $f(Si, Aj) = V_{SiAj}$. Số mức giá trị *l* của mỗi thuộc tính A_j có thể khác nhau (như ở *Bảng 4*) tùy vào phương pháp định lượng hóa sao cho đảm bảo độ mịn và tiệm cận với các mức đo của nhà sản xuất nút cảm biến.

Bảng 3. IS tại thời điểm bắt đầu tiền xử lý

U	Q (tập thuộc tính)			
	A_1	A_2	.....	A_m
S_1	$V_{A1.S1}$	$V_{A2.S1}$	.....	$V_{Am.S1}$
S_2	$V_{A1.S2}$	$V_{A2.S2}$	.....	$V_{Am.S2}$
S_2	$V_{A1.S2}$	$V_{A2.S2}$	.....	$V_{Am.S2}$
....	.....	.....	.....	.....
S_n	$V_{A1.Sn}$	$V_{A2.Sn}$	.....	$V_{Am.Sn}$

Bảng 4. Giá trị các thuộc tính A_j

V	Q (tập thuộc tính)			
	A_1	A_2	.....	A_m
X_1	$X_{1.A1}$	$X_{1.A2}$	.....	$X_{1.Am}$
X_2	$X_{2.A1}$	$X_{2.A2}$	.....	$X_{2.Am}$
....	.....	.....	.....	.....
X_l	$X_{l.A1}$	.....	.....	.....
		.....	.....	$X_{l.Am}$
		$X_{l.A2}$	.....	

1. Dữ liệu thiếu

Dữ liệu thu thập được từ các nút cảm biến khi truyền đến CH có thể không đầy đủ, nghĩa là CH không nhận đủ dữ liệu đo về một hoặc nhiều tham số đo từ một hoặc nhiều nút trong nhóm gửi về để làm dữ kiện cho quá trình DF. Tình huống để mất dữ liệu có thể là: Lúc cần cảm nhận hoặc truyền dữ liệu đến đích thì nút cảm biến đang trạng thái ngủ, lúc đang truyền dữ liệu đến CH thì nút cảm biến hết năng lượng,…

Dữ liệu bị thiếu có thể là toàn bộ kết quả đo mà nút cảm biến ghi nhận từ mục tiêu trong cả chu kỳ *T* hoặc trong 1 khung tin F_k hoặc 1 phần của khung tin (là 1 hoặc nhiều tham số P_j nào đó trong F_k nào đó) hoặc tất cả các yếu tố trên. Không mất tính tổng quát, có thể xem tại thời điểm CH đóng gói xong để tiền xử lý, dữ liệu đo của nút cảm biến S_i $(1 \le i \le n)$ với tham số đo A_j $(1 \le j \le m)$ bị thiếu, ký hiệu $f(S_i, A_j) = \varnothing$ (4)

Mạng cảm biến không dây sử dụng giao thức IEEE 802.15.4 sẽ điều khiển việc lấy dữ liệu theo chu kỳ thức-ngủ *(active-sleep)* nên dữ liệu CH thu được từ nút cảm biến có tính rời rạc, $f(Si, Aj)$ có thể được tính thông qua xác suất, các giá trị có tính ngẫu nhiên trong miền giá trị đo X_l của thuộc tính A_j. Chúng tôi áp dụng Entropy Shannon [7] để tính xác suất xuất hiện của *l* khả năng (giá trị) của thuộc tính A_j tương ứng. Gọi $Pr(X_t)$ là xác suất xuất hiện giá trị X_t $(1 \le t \le l)$ của thuộc tính A_j, Entropy Shannon *(ES)* của tập *U* (nút cảm biến) đối với A_j được tính như sau:

$$ES(U) = -\sum_{t=1}^{l} Pr(X_t)\ \log Pr(X_t) \quad (5)$$

Gán $f(S_i, A_j).\varnothing = Max\ Pr(X_t)$ (6)

Trong đó $f(S_i, A_j).\varnothing$ là dữ liệu đo bị thiếu của nút cảm biến S_i về thuộc tính A_j, $Max\ Pr(X_t)$ là giá trị X_t mà khả năng $f(S_i, A_j)$ nhận được nhất (hay $Pr(X_t)$ lớn nhất). Biến ngẫu nhiên X_t có thể nhận *l* mức, xác suất $1/l$. Thuộc tính A_j có thể xem là biến ngẫu nhiên với xác suất luôn bằng 1.

2. Dữ liệu nhiễu (yếu)

Do các nút cảm biến truyền dữ liệu bằng sóng vô tuyến đến CH nên tín hiệu bị yếu (về cường độ) bởi các yếu tố gây nhiễu ở trong môi trường. Trong bài báo này, chúng tôi giả sử đã phát hiện được nhiễu, tức là đã xác định được kết quả đo thuộc tính A_j của nút cảm biến S_i đã bị nhiễu, cần phải xử lý.

Gọi λ là ngưỡng giá trị đo của thuộc tính A_j $(1 \le j \le m)$. Dữ liệu đo của S_i gọi là nhiễu (yếu) nếu $f(S_i, A_j) \le \lambda$. Gọi $f_{noisy}(S_i, A_j)$ là giá trị nhiễu của S_i khi đã đo tham số A_j, $P.f_{noisy}(S_i, A_j)$ là xác suất $f_{noisy}(S_i, A_j)$ đúng với $f(S_i, A_j)$ (là giá trị không nhiễu), khi đó $P.f_{noisy}(S_i, A_j) \le 1$ và sai số $\delta = \frac{f_{noisy}(Si, Aj)}{f(Si, Aj)} \le 1$ (7)

Nếu $f_{noisy}(S_i, A_j)$ có $P.f_{noisy}(S_i, A_j) \ge 0.5$ khả năng $f_{noisy}(S_i, A_j)$ là tín hiệu nhiễu lớn hơn mức trung bình. Giả sử giá trị nhiễu $f_{noisy}(S_i, A_j)$ sau khi đã xử lý là $f_{fix}(S_i, A_j)$ với $X_{1.Aj} \le f_{fix}(S_i, A_j) \le X_{l.Aj}$. Để đảm bảo tính toàn vẹn *(completeness)* của dữ liệu cảm nhận và giảm nguy cơ sai số tích lũy khi sử dụng dữ liệu này làm đầu vào quá trình DF, chúng tôi đề xuất mối quan hệ giữa 2 giá trị này như sau: $f_{fix}(S_i, A_j) = f_{noisy}(S_i, A_j)/2$ (8)

Với giá trị ngưỡng λ, tùy theo từng thuộc tính để lựa chọn giá trị ngưỡng λ phù hợp và sai số δ tương ứng. Ví dụ một công thức tính ngưỡng ở [3] là trung bình cộng của l mức giá trị đo thành phần của thuộc tính tương ứng trong điều kiện tiêu chuẩn thiết kế, ví dụ ngưỡng giá trị của thuộc tính A_j là $\lambda = \left(\sum_{t=1}^{l} X_{t\,Aj} \right) / l$ (9)

3. Giải thuật xử lý dữ liệu thiếu, nhiễu

```
Set n = num_nodes; set m = num_condi_attrib
1   For {set i 1} {$i <= $n } {incr i}
2     For {set j 1} {$j <= $m} {incr j}
3       if V_Si.Aj = ∅ then
4             P. V_Si.Aj (X_l, A_j)
5             Select [Max (P. V_Si.Aj (X_l, A_j)) and (Max (P. V_Si.Aj (X_l, A_j)) ≥ λ)]
6             Set V_Si.Aj = Max (P. V_Si.Aj (X_l, A_j))  # Gán vào giá trị ∅
7             Else if Select [Max (P. V_Si.Aj (X_l, A_j)) and (Max (P. V_Si.Aj (X_l, A_j)) < λ)] then
8             Set V_Si.Aj = Max (P. V_Si.Aj (X_l, A_j)) and Set V_Si.Aj = Vnoisy Si.Aj
9       if [(V_Si.Aj = Vnoisy Si.Aj) and (P.f_noisy(S_i, A_j) ≥ 0.5)] then
10            V_Si.Aj = Vnoisy Si.Aj / 2
11            Else Set V_Si.Aj = Vnoisy Si.Aj
12  Return V_Si.Aj
```

Đặt số nút cảm biến trong cụm là n, số thuộc tính là m. Dòng 1, 2 để lọc hết các giá trị đo của nút cảm biến S_i về tham số A_j trong bảng. Dòng 3, nếu giá trị đo của S_i về A_j bị nhiễu thì (Dòng 4) tính xác suất có điều kiện Entropy Shannon của l khả năng giá trị X_t $(1 \le t \le l)$ đối với tham số A_j và (Dòng 5) lựa chọn giá trị X_t có xác suất đúng với $V_{Si.Aj}$ nhất đồng thời bằng mức ngưỡng λ trở lên thì (Dòng 6) gán giá trị này vào $V_{Si.Aj}$. Dòng 7 là trường hợp ngược lại của $X_t < \lambda$ thì (Dòng 8) gán giá trị đó vào $V_{Si.Aj}$ đồng thời đánh dấu là tín hiệu nhiễu để được xử lý ở bước sau. Dòng 9, nếu $V_{Si.Aj}$ là nhiễu với xác suất nhiễu là đúng (là nhiễu thật) $P.f_{noisy}(S_i, A_j) \ge 0.5$ thì (Dòng 10) giảm giá trị nhiễu này xuống một nửa để giảm nguy cơ sai số tích lũy đồng thời đảm bảo tính toàn vẹn của dữ liệu. Các giá trị nhiễu khác có xác suất đúng dưới 0.5 thì được giữ nguyên (ở dòng 11). Dòng 12, sau 2 lượt xử lý dữ liệu thiếu (Dòng 3 đến 8) và xử lý dữ liệu nhiễu (Dòng 9 đến 11), các $V_{Si.Aj}$ của bảng được hoàn thành làm đầu vào để xử lý dữ liệu dư thừa ở bước sau.

B. Xử lý dữ liệu dư thừa

Lý thuyết tập thô được đề xuất và chứng minh là có thể áp dụng để tổng hợp dữ liệu nhiều cảm biến [2]. Theo *Hình 2*, đầu ra của quy trình xử lý tín hiệu thiếu và nhiễu là bảng dữ kiện đầy đủ gồm n hàng, m cột hay là ma trận cỡ *(n x m)*. Bảng này sẽ là đầu vào của khối xử lý dữ liệu dư thừa. Mục đích của xử lý dữ liệu dư thừa là rút gọn bảng cụ thể là phải tìm được thuộc tính lõi, tập thuộc tính rút gọn.

Thuộc tính lõi là hợp tất cả các tập một phần tử trong ma trận phân biệt (hay ma trận khả phân) [5].

$$Core(A) = \{a \in A: c_{ur} = \{a\}, (u, r = 1, 2, ..., n)\} \quad (10)$$

Ma trận phân biệt (discernibility matrix): Là một ma trận đối xứng cỡ *(n x n)*, ký hiệu *M(IS)*, giá trị của các phần tử c_{ur} của ma trận *M(IS)* được định nghĩa như sau [5]:

$$(c_{ur}) = \{a \in A: a(S_i) \ne a(S_j)\} \text{ đối với } \forall i, j = 1, 2, ..., n \quad (11)$$

Tập thuộc tính rút gọn: Tập A' được gọi là tập thuộc tính rút gọn của A nếu $A' \subseteq A$ và A' là tập con cực tiểu của A sao cho $A' \cap c \ne \varnothing$ với $\forall c \in M(IS)$ [5].

Hiện nay, các nghiên cứu (quốc tế và Việt Nam) đã đưa ra 5 phương pháp để rút gọn thuộc tính như sau [6]:

- Dựa trên miền dương;
- Sử dụng các phép toán trong đại số quan hệ;
- Sử dụng ma trận phân biệt;
- Sử dụng các độ đo trong tính toán hạt;
- Sử dụng entropy thông tin.

Trong bài báo này, chúng tôi sẽ áp dụng phương pháp ma trận phân biệt. Từ định nghĩa về ma trận phân biệt ở *(11)*, chúng tôi xây dựng hàm phân biệt theo [5]. *Hàm phân biệt Fis* được tạo nên từ ma trận phân biệt *M(IS)* là một hàm Boolean có *m* biến Boolean (tương đương với *m* thuộc tính $A_1, A_2, ..., A_m$) được xây dựng dưới dạng chuẩn tắc hội (hội của các tuyển sơ cấp) như sau [5]:

$$Fis\ (A^*_1, A^*_2, \ldots, A^*_m) = \otimes \{ \oplus c_{ur} \mid 1 \le u \le r \le n, c_{ur} \neq \varnothing \}, \text{ trong đó } c^*_{ur} = \{A^* \mid A \in c_{ur}\} \quad (12)$$

Sau khi rút gọn hàm Boolean ở (*12)*, tập các đơn thức của *Fsi* xác định tập rút gọn của bảng dữ liệu.

C. Ví dụ minh họa

Giả sử một mạng cảm biến có 5 nút cảm biến $S_1 \rightarrow S_5$ với 4 thuộc tính $A_1 \rightarrow A_4$ với kết quả đo được ở *Bảng 5*:

Bảng 5. Dữ liệu đo của mạng cảm biến

Nút cảm biến	Thuộc tính			
	A_1	A_2	A_3	A_4
S_1	∅	5	~ 2	4
S_2	3	4	∅	4
S_3	3	3	3	4
S_4	4	∅	4	4
S_5	∅	~ 2	4	∅

- Tập các dữ liệu bị thiếu, tính theo *(4)* gồm : $\{(S_1, A_1), (S_2, A_3), (S_4, A_2), (S_5, A_1), (S_5, A_4)\}$
- Tập các dữ liệu bị nhiễu gồm: $\{(S_1, A_3), (S_5, A_2)\}$
- Tập giá trị đo của các thuộc tính, được xác định theo *(3* và *Bảng 4*:

  V_{A1} = {1, 2, 3, 4, 5, 6} = {rất nhỏ, nhỏ, trung bình, mạnh, rất mạnh, năng lượng gốc}, $l = 6$

  V_{A2} = {1, 2, 3, 4, 5} = {rất xa, xa, trung bình, gần, rất gần}, $l = 5$

  V_{A3} = {1, 2, 3, 4, 5} = {còn rất nhiều, còn nhiều, còn gần 1/2, còn ít, còn rất ít}, $l = 5$

  V_{A4} = {1, 2, 3, 4, 5} = {rất lớn, lớn, trung bình, ít, rất ít}, $l = 5$
- Ngưỡng các giá trị thuộc tính được tính theo *(9)*: $\lambda.A_1 = 3{,}5$; $\lambda.A_2 = \lambda.A_3 = \lambda.A_4 = 3$
- Xác suất nhận được giá trị đo theo thuộc tính tương ứng của các nút cảm biến được tính theo *(5)*, giả sử kết quả làm tròn số đối với dữ liệu đo bị thiếu ở *Bảng 6* và dữ liệu đo bị nhiễu ở *Bảng 7*:

Bảng 6. Xác suất khi dữ liệu đo bị thiếu

Giá trị đo	Xác suất nhận giá trị đo theo thuộc tính				
	(S1,A_1)	(S2,A3)	(S4,A2)	(S5,A_1)	(S5,A4)
1	0,1	0,1	0,2	0,1	0,1
2	0,1	0,1	0,1	0,2	0,1
3	0,2	0,2	0,1	0,1	0,3
4	**0,4**	**0,5**	**0,4**	0,2	0,1
5	0,1	0,1	0,2	0,1	**0,4**
6	0,1			**0,3**	

Bảng 7. Xác suất khi dữ liệu đo bị nhiễu

Giá trị đo	Xác suất nhận giá trị đo theo thuộc tính	
	(S1,A3)	(S5,A2)
1	0,1	0,1
2	**0,6**	**0,5**
3	0,1	0,1
4	0,1	0,1
5	0,1	0,2
6		

Sau khi xử lý dữ liệu thiếu, nhiễu theo thuật toán, được kết quả như ở *Bảng 8*:

Bảng 8. Dữ liệu đã xử lý thiếu, nhiễu

Nút cảm biến	A_1	A_2	A_3	A_4
S_1	4	5	1	4
S_2	3	4	4	4
S_3	3	3	3	4
S_4	4	4	4	4
S_5	6	1	4	5

Bảng 9. Ma trận phân biệt

Nút cảm biến	S_1	S_2	S_3	S_4	S_5
S_1	∅				
S_2	A_1,A_2,A_3	∅			
S_3	A_1,A_2,A_3	A_2,A_3	∅		
S_4	A_2,A_3	A_1	A_1,A_2,A_3	∅	
S_5	A_1,A_2,A_3,A_4	A_1,A_2,A_4	A_1,A_2,A_3,A_4	A_1,A_2,A_4	∅

- Xây dựng Ma trận phân biệt: Từ *Bảng 8*, áp dụng (*11)*, được ma trận phân biệt ở *Bảng 9*.
- Hàm phân biệt: Từ *Bảng 9* và (*12)* để xây dựng hàm phân biệt F_{Si}. Đây là một hàm Boolean dạng chuẩn tắc hội (hội của các tuyển sơ cấp) có 4 biến Boolean như sau:

$$F_{IS}(A_1,A_2,A_3,A_4) = (A_1 \vee A_2 \vee A_3) \wedge (A_1 \vee A_2 \vee A_3) \wedge (A_2 \vee A_3) \wedge (A_2 \vee A_3) \wedge A_1 \wedge (A_1 \vee A_2 \vee A_3) \wedge$$
$$\wedge (A_1 \vee A_2 \vee A_3 \vee A_4) \wedge (A_1 \vee A_2 \vee A_4) \wedge (A_1 \vee A_2 \vee A_3 \vee A_4) \wedge (A_1 \vee A_2 \vee A_4). \quad (13)$$

• Tập thuộc tính rút gọn: Kết quả của việc rút gọn hàm Boolean ở (*13)* sẽ là tập thuộc tính rút gọn. Trong bài báo này, chúng tôi sử dụng các phép toán trong đại số quan hệ, cụ thể có hai cách: phương pháp đại số hoặc bìa Các-nô (*Carnaugh*), kết quả rút gọn như sau: $F_{IS}(A_1,A_2,A_3,A_4) = (A_1 \wedge A_2) \vee (A_1 \wedge A_3) \vee (A_1 \wedge A_3 \wedge A_4)$. Như vậy có 3 tập thuộc tính rút gọn: $RG_1=\{A_1, A_2\}$; $RG_2=\{A_1, A_3\}$; $RG_3=\{A_1, A_3, A_4\}$. (14)

• Tập thuộc tính lõi: Theo (*10)* và *Bảng 9* (hoặc (*14*)) tính được Core(A) = $\{A_1\}$.

Bảng 10. Các phương án dữ kiện đã được rút gọn làm đầu vào để CH tổng hợp dữ liệu

(a). Dữ liệu RG1

Sensor node	A_1	A_2
S_1	4	5
S_2	3	4
S_3	3	3
S_4	4	4
S_5	6	1

(b). Dữ liệu RG2

Sensor node	A_1	A_3
S_1	4	1
S_2	3	4
S_3	3	3
S_4	4	4
S_5	6	4

(c). Dữ liệu RG3

Sensor node	A_1	A_3	A_4
S_1	4	1	4
S_2	3	4	4
S_3	3	3	4
S_4	4	4	4
S_5	6	4	5

Đối với giải pháp tổng hợp dữ liệu nhiều cảm biến của mạng cảm biến không dây bằng lý thuyết tập thô nói chung và các phương pháp khác sử dụng bảng dữ liệu cảm biến làm cơ sở dữ kiện đầu vào để CH tiến hành tổng hợp thì việc tìm tập thuộc tính rút gọn, thuộc tính lõi là một trong những bước thực hiện quan trọng. Giải pháp DP-DF đề xuất sử dụng dữ liệu của các bảng rút gọn ở *Bảng 10* làm đầu vào để DF. Tùy mục đích và giải thuật DF để lựa chọn các bảng này đảm bảo cân bằng giữa tính toán và bảo toàn dữ liệu cảm nhận của nút cảm biến trong cụm.

IV.KẾT LUẬN

Phương pháp DP-DF đã đề xuất được giải pháp tiền xử lý dữ liệu cảm nhận của các nút trong cụm khi gửi về nút cụm trưởng CH để tổng hợp dữ liệu. Giải pháp bao gồm *xử lý dữ liệu thiếu và dữ liệu nhiễu bằng xác suất Entropy Shannon* và *xử lý dữ liệu dư thừa bằng lý thuyết tập thô.*

Hướng nghiên cứu tiếp theo: Phát hiện nhiễu tín hiệu bằng xác suất; nghiên cứu sử dụng các bảng dữ liệu đã rút gọn nhằm đảm bảo tối ưu trong tổng hợp dữ liệu nhiều cảm biến.

V. TÀI LIỆU THAM KHẢO

[1] W. Heinzelman, A.P. Chandrakasan and H. Balakrishnan, *"Energy-Efficient Communication Protocol for Wireless Microsensor Networks"*, IEEE Proceedings of the Hawaii International Conference on System Sciences, January 4-7, 2000, Maui, Hawaii.

[2] Dương Viết Huy, Nguyễn Duy Tân, Hồ Đức Ái, Nguyễn Đình Việt, "*Tiếp cận phương pháp tổng hợp dữ liệu nhiều cảm biến trong mạng cảm biến không dây bằng lý thuyết tập thô",* Proceedings of the 7th National Conference on Fundamental and Applied IT Research (FAIR'7), 2014, pp. 668-677.

[3] Dương Viết Huy, Nguyễn Đình Việt, *"Đề xuất giải pháp tổng hợp dữ liệu nhiều cảm biến trong mạng cảm biến không dây"*, Kỷ yếu Hội thảo Quốc gia lần thứ XVII (Hội thảo @2014), trang 50-55.

[4] Jiawei Han, Micheline Kamber, *"Data Mining: Concepts and Techniques"*, Second Edition, Morgan Kaufmann Publishers, 2006, pp. 47-80.

[5] Zdzisaw Pawlak, *"Rough sets"*, International Journal of Computer and Information Sciences, 11, 341-356, 1982.

[6] Nguyễn Long Giang, *"Luận án Tiến sỹ toán học: Nghiên cứu một số phương pháp khai phá dữ liệu theo tiếp cận lý thuyết tập thô"*, 2012, trang 2-3.

[7] Annick Lesne, *"Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics"*, Mathematical Structures in Computer Science, Volume 24 , Special Issue 03, June 2014.

PROPOSED SOLUTION TO PREPROCESSING FOR MULTI-SENSOR DATA FUSION IN WIRELESS SENSOR NETWORK

Duong Viet Huy, Nguyen Dinh Viet

ABSTRACT - When using multiple sensor nodes for monitoring (measuring) multiple-parameters of the target, the measuring data are often the same information. This redundant data is sent to base station (BS) causes the waste of energy of sensor nodes and the risk of congestion. One of the factors that affect the quality output data fusion is quality of input data to cluster head node (CH) from the nodes in the cluster. Because sensor node transceiver data by electromagnetic wave signal, so there are many factors that affect measuring target as noise, loss of data... If CH uses these measuring results as data input for data fusion, it may not measure every change of target. This paper, we proposed DP-DF solution.to pre-process, remove the raw data and remain data that have more valuable knowledge to data fusion.