O Uso do Algoritmo de Agrupamento Hierárquico Divisivo DIANA em uma Rede de Sensores Sem Fio Aplicada à Agricultura

Paulo R. Nietto, Hugo V. Sampaio

Faculdade Campo Limpo Paulista – FACCAMP
Campo Limpo Paulista – SP
{pnietto,hvazsampaio}@cc.faccamp.br

Abstract. *Wireless Sensor Networks (WSF) use sensor nodes to obtain information about the characteristics of the environment. The information collected helps on decisions to activate actuators such as, the irrigation frequency or control the air temperature inside a greenhouse. The data can also be used as input for divisive hierarchical clustering algorithms, and based on the data clustering, it is possible to derive conclusions. This article describes the main steps to achieve the proposed objectives.*

Resumo. *Uma rede de sensores sem fio (RSSF) usa nós sensores para obter informações sobre as características do domínio de sua aplicação. Tais informações auxiliam na decisão de atuadores, como por exemplo definir a frequência de irrigação ou controlar a temperatura do ar de uma estufa. Os dados captados por RSSF podem também servir de entrada a algoritmos de agrupamento hierárquicos divisivos para que, com base no agrupamento de tais dados, seja possível derivar conclusões sobre eles. O artigo descreve uma proposta de trabalho e os principais passos para atingir os objetivos propostos.*

1. Introdução

O avanço tecnológico de rede de sensores sem fio (RSSF) está proporcionando inovações em várias áreas de aplicação, tais como, médica, agrícola, controle de tráfego, controle ambiental entre outras. Na área de agricultura a RSSF pode ser usada para obter as informações necessárias que uma planta necessita para crescer, como temperatura e umidade do ar, umidade da terra, luminosidade, entre outros. As variações dessas informações são relativamente lentas, assim, os sensores sem fio, alimentados por bateria, podem coletar e enviar essas informações esporadicamente, de tal modo que um sistema de monitoramento agrícola poderia informar tais dados a uma equipe de plantão ou automaticamente tomar providências como, por ex., regar, ligar sistema de ventilação, deixar o ambiente mais escuro, etc.

Em várias taxonomias propostas na literatura, com o objetivo de organizar os algoritmos de agrupamento (ver Xu *et al*. (2007), Jain (2010), Theodoridis & Koutroumbas (2009)), uma categoria sempre presente é a dos chamados algoritmos

hierárquicos que, via de regra, produzem um conjunto de agrupamentos aninhados, organizados como uma árvore hierárquica, que pode ser visualizada como um dendrograma. Os algoritmos hierárquicos, por sua vez, se subdividem em dois grupos: (1) hierárquicos aglomerativos, que se caracterizam por iniciar o processo com um agrupamento em que cada um dos dados é considerado um grupo do agrupamento e (2) hierárquicos divisivos, que se caracterizam por iniciar o processo com um agrupamento tendo apenas um grupo, aquele com todos os dados fornecidos e que, a cada passo, divide um dos grupos do agrupamento, até que cada grupo contenha um dado (ou então, existam apenas k grupos, em que k é fornecido como parâmetro), em uma abordagem caracterizada como *top-down*.

Este artigo descreve os passos já realizados relativos a um projeto de pesquisa voltado à investigação do uso do algoritmo de agrupamento hierárquico divisivo DIANA que tem por objetivo agrupar os dados captados por RSSF para que seja possível identificar similaridades e diferenças nos dados e, como consequência, seja possível derivar conclusões sobre eles. A Seção 2 contextualiza as RSSF. Na Seção 3 é apresentada uma descrição geral do algoritmo de agrupamento hierárquico divisivo DIANA para, então, apresentar seu pseudocódigo. A Seção 4 apresenta o experimento preliminar. Por fim, a Seção 5 descreve os próximos passos para a continuação e finalização do projeto de pesquisa.

2. Rede de Sensores Sem Fio

As redes de sensores sem fio (RSSF) são constituídas por um conjunto de dispositivos interligados denominados nós de rede. Um nó é composto por microprocessador, memória, rádio e fonte de energia. Uma RSSF possui um nó coordenador que cria e gerencia a RSSF e, nós sensores (um nó de rede com 1 ou mais sensores) para coleta de dados do domínio ao qual estão incorporados [Yick *et al.* 2008, Culler *et al.* 2004].

Este artigo apresenta uma RSSF desenvolvida com padrão Zigbee [Zigbee 2016], contendo um nó coordenador e três nós sensores. Cada nó utiliza uma antena Xbee Zigbee [Xbee Zigbee 2016], conectada a um Arduíno Uno [Arduino Uno 2016]. A RSSF foi desenvolvida em ambiente controlado para coletar dados ambientais de luminosidade, temperatura/umidade do ar e umidade da terra, através de nós conectados a sensores. Para cada nó sensor foi inserido um algoritmo no microcontrolador para captura e envio de dados. Cada sensor é conectado ao Arduino Uno através de uma porta analógica e a captura de dados é realizada através da leitura dessas portas.

O rádio Xbee Zigbee utilizado neste projeto possui uma limitação de 14 nós filhos e alcance de sinal em 60 metros para ambientes fechados. As antenas são configuradas utilizando o software XCTU [XCTU 2016], fornecido no site da Digi [Digi 2016]. Através do XCTU podemos definir o modo de funcionamento das antenas (Coordenador, Roteador ou *End Device*), PANID (*Personal Area Network Id*), endereço de destino, entre outros.

O Arduíno Uno, utilizado em cada nó, é constituído de uma placa com microcontrolador ATmega328P que possui clock de 16 MHz, 32KB de memória flash, 14 portas digitais para *input*/*output*, 6 portas analógicas e 1 porta USB. Uma antena Xbee Zigbee é conectada em cada Arduino Uno para transmissão dos dados captados.

Os sensores de temperatura e umidade do ar, do tipo DHT11 [DHT11 2016], possui uma biblioteca para Arduíno, onde os valores capturados indicam a umidade do ar em porcentagem (%), e a temperatura do ar em graus centígrados (Cº). A precisão da temperatura é de ±2 graus, e a precisão da umidade do ar é de ±5%.

O nó sensor de luminosidade possui um LDR (*Light Dependent Resistor*), onde os valores capturados são entre 0 e 1023 e tais valores foram convertidos em uma escala de 0 a 9, onde 0 indica menor incidência de luz e 9 maior incidência de luz. O sensor de detecção de variações de umidade do solo captura valores entre 0 e 1023, onde tais valores foram convertidos em uma escala de 0 a 9, sendo 0 completamente molhado e 9 completamente seco.

Os nós sensores foram configurados para capturar os dados dos sensores com intervalo de um minuto por um período de 24 horas. Os dados capturados foram encaminhados para o nó coordenador, que transferiu as informações para um computador usando a porta Serial/USB. Cada nó sensor capturou 1440 dados, totalizando 5760 dados.

3. O Algoritmo DIANA

O algoritmo *DIvisive ANAlysis* (DIANA), é um algoritmo divisivo baseado na proposta de Macnaughton-Smith *et al.,* (1964), que possui a complexidade $O(N^2 \log N)$. Como descrito em Kaufman & Rousseeuw (1990), é uma proposta de um procedimento para algoritmos hierárquicos divisivos, em que algumas das funções empregadas podem ser customizadas, na dependência da aplicação considerada. Devido à possibilidade de customização, tal esquema pode dar origem a diferentes "instanciações", muitas vezes consideradas novos algoritmos.

Segue uma descrição informal do algoritmo. No processo iterativo conduzido pelo DIANA, o t-ésimo agrupamento produzido tem t grupos em que t = 1, ..., N, onde N é a quantidade de dados.

Considere que G seja um grupo já formado e que o objetivo seja dividi-lo em dois grupos de tal maneira que os dois grupos resultantes possuam a maior dissimilaridade possível entre eles. Inicialmente o algoritmo busca identificar um dado em G cuja dissimilaridade média com relação aos dados restantes seja máxima. O dado com dissimilaridade máxima é retirado de G e inserido em um novo grupo criado nesse momento, chamado tempG. Na sequência, para cada dado $x \in G$, é calculada a média dos valores de dissimilaridade de x, com todos os demais dados de G, i.e., média dos valores diss(x,y), $y \in G$. De maneira análoga, é calculada a média dos valores de dissimilaridade de x com os dados pertencentes a tempG, ou seja, a média dos valores diss(x,z) $z \notin G$ (i.e, $z \in$ tempG).

Se para cada $x \in G$:

- D(x) = (média diss(x,y), $y \in G$, $x \neq y$) − (média diss(x,z), $z \in$ tempG) for negativa, então tempG não receberá mais nenhum dado de G.

- Caso contrário, o dado $x \in G$ que produzir o valor máximo para a diferença D(x) = (média diss(x,y), $y \in G$, $x \neq y$) – (média diss(x,z) $z \in$ tempG) é escolhido e retirado de G e, então, inserido em tempG.

O procedimento é repetido até que cada dado seja o único em um grupo ou, então até que um critério de parada seja satisfeito. O Algoritmo 3.1 descreve um pseudocódigo do algoritmo DIANA.

```
procedure DIANA (X, AGt)
Input: X = {P1, P2,...,PN}  %N dados a serem agrupados
Output: AGt

1. begin
2. t ← 1
3. AG1 ← {X}                % agrupamento inicial
4. G ← X
5. repeat
6.   maxDiss ← maiorDissimilaridade(G)    % encontrar elemento mais dissimilar
7.    G ← G – {maxDiss}
8.    tempG ←{maxDiss}
9.   repeat
10.     Para cada dado x pertencente a G encontre:
12.        D(x) = (média diss(x,y), x ∈ G e x≠y] – (média diss(x,z) z ∈ tempG)
11.        Encontre o dado x para qual a diferença D(x) é a maior.
12.        if D(x) > 0 then
13.           begin
14.              tempG ← tempG ∪ {x}
15.              G ← (G – {x})
16.           end
17.  until todos valores D(x) ≤ 0
18.  t ← t+1
19. Gt ← tempG
20. AGt ← (AGt-1 – {G}) ∪ {G, Gt}
21. G ← grupoMaiorDiametro(AGt)          % encontrar grupo com maior diâmetro
22. until cada grupo contenha um único dado
23.return AGt
24.end procedure
```

Algoritmo 3.1. Pseudocódigo do algoritmo DIANA.

4. Experimento Preliminar

O experimento preliminar foi realizado utilizando a RSSF descrita na Seção 2 e o algoritmo DIANA, descrito na Seção 3. O algoritmo DIANA foi adotado nesse trabalho com objetivo de verificar sua eficácia em agrupar dados advindos do domínio da agricultura. O algoritmo K-Means é utilizado nesse experimento para comparação de resultados, por ser o algoritmo mais conhecido e utilizado para a tarefa de agrupamento de dados e, também, para apoio aos diversos outros algoritmos que possuem alto custo computacional. O conjunto de dados utilizado consiste de 1440 dados descritos por 4

atributos, onde cada atributo corresponde ao valor capturado por um dos sensores no mesmo minuto.

Os resultados obtidos foram avaliados utilizando dois índices de validação interna: o índice de Dunn [Dunn 1974], apresentado na Eq. (4.1) e o índice Silhouette [Rowsseew 1987], apresentado na Eq. (4.2). A Tabela 4.1 descreve a notação utilizada nos índices de validação. Índices de validação interna são métodos adequados para a avaliação quantitativa do resultado de um agrupamento, uma vez que tal avaliação do agrupamento resultante de um algoritmo utiliza somente quantidades e características inerentes ao conjunto de padrões.

Tabela 4.1 Notação utilizada nos índices de validação**.**

Notação	Significado
P_i	I-ésimo padrão
G_i	I-ésimo grupo
NG	Número de grupos
$d(P_i,P_j)$	Distância entre dois padrões distintos
$\|G_i\|$	Quantidade de dados do i-ésimo grupo

$$\text{Dunn} = \min_{i=1,\dots,NG}\left\{\min_{j=i+1,\dots,NG}\left(\frac{d(G_i,G_j)}{\max_{K=1,\dots,NG}\text{diam}(G_K)}\right)\right\}, \qquad (4.1)$$

$$d(G_i,G_j) = \min_{P_i\in G_i, P_j\in G_j}\{d(P_i,P_j)\}; \quad \text{diam}(G_i) = \max_{P_i,P_j\in G_i}\{d(P_i,P_j)\}$$

$$\text{Silhouette} = \frac{1}{NG}\sum_i\left\{\frac{1}{|G_i|}\sum_{P_i\in G_i}\frac{b(P_i)-a(P_i)}{\max[b(P_i),a(P_i)]}\right\}, \qquad (4.2)$$

$$a(P_i) = \frac{1}{|G_i|-1}\sum_{P_j\in G_i} d(P_i,P_j); \quad b(P_i) = \min_{j,j\neq i}\left[\frac{1}{|G_j|}\sum_{P_j\in G_j} d(P_i,P_j)\right]$$

Devido à versão do algoritmo K-Means utilizada ser sensível à escolha dos centroides iniciais (que são escolhidos de forma aleatória), o algoritmo K-Means é executado cinco vezes e, tanto a média dos cinco resultados quanto o melhor dos cinco resultados obtidos são utilizados para comparação.

4.1 Resultado do Agrupamento dos Dados

A Tabela 4.2 ilustra o resultado dos índices de validação Dunn e Silhouette nos agrupamentos resultantes dos algoritmos DIANA e K-Means com o valor do parâmetro de número de grupos 2 (*ie.*, K=2). Apesar do algoritmo DIANA não requerer o número de grupos como parâmetro, esse parâmetro pode ser utilizado como critério de parada.

Tabela 4.2 Resultados dos agrupamentos no conjunto de padrões captados por RSSF com K=2.

Algoritmo	Dunn	Silhouette
DIANA	0,58	0,51
K-Means (melhor execução)	0,08	0,59
K-Means (média das execuções)	0,08	0,59

O algoritmo K-Means induziu o mesmo agrupamento em suas 5 execuções. O índice de Dunn indica que o agrupamento induzido pelo algoritmo DIANA é superior ao induzido pelo algoritmo K-Means. Entretanto, segundo o índice Silhouette, o algoritmo K-Means apresenta performance superior ao algoritmo DIANA. Se considerar que o índice de Dunn apresentou um valor 0,5 superior para agrupamento resultante DIANA e o índice Silhouette apresentou um valor 0,08 acima para o agrupamento resultante do K-Means, pode-se concluir que nesse experimento, o algoritmo DIANA produziu um agrupamento mais representativo em relação ao agrupamento produzido pelo algoritmo K-Means.

5. Comentários Finais e Próximas Etapas

O trabalho até então desenvolvido, que envolveu testes de captura dos dados através da RSSF e o uso de tais dados como entrada ao algoritmo DIANA, continuará com o desenvolvimento e aplicação de uma RSSF na área de agricultura para que os dados captados pela RSSF sejam utilizados como entrada ao algoritmo DIANA e, através dos agrupamentos resultantes do algoritmo, seja possível identificar similaridades e diferenças nos dados e, como consequência, derivar conclusões.

Referências

Arduíno Uno (2016). Disponível em <www.arduino.cc/en/Main/ArduinoBoardUno> [Acessado em 16/06/2016].

Culler, D., Estrin, D. and Srivastava, M. (2004) "Overview of Sensor Networks". *IEEE Computer Magazine*, v. 37, no. 8, pp. 41–48.

DHT11 (2016). Disponível em <http://www.micropik.com/PDF/dht11.pdf> [Acessado em 05/07/2016].

Digi (2016). Disponível em <http://www.digi.com/> [Acessado em 28/06/2016].

Dunn, J. (1974) "Well Separated Clusters and Optimal Fuzzy Partitions". *Journal of Cybernetics*, v. 4, pp. 95–104.

Jain, A. K. (2010) "Data clustering: 50 years beyond K-means". Pattern Recognition Letters, v. 31, pp. 651–666.

Kaufman, L. & Rousseeuw, P.J. (1990) "Finding Groups in Data: An Introduction to Cluster Analysis". USA: John Wiley & Sons, Inc.

Rowsseew, P. (1987) "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics*, v. 20, pp. 53–65

Theodoridis, S. & Koutroumbas, K. (2009) "Pattern Recognition", 4th edition. USA: Elsevier.

Xbee Zigbee (2016). Disponível em <http://www.digi.com/products/xbee-rf-solutions/rf-modules/xbee-zigbee> [Acessado em 16/06/2016].

Xu, H., Xu, D., Lin, E. (2007) “An applicable hierarchical clustering algorithm for content-based image retrieval”, In: Proc. of The International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, MIRAGE’07, pp.82–92.

Yick, J., Mukherjee, B. and Ghosal, D. (2008) “Wireless Sensor Network Survey”. *Computer Networks*, v.52, no. 12, pp. 2292–2330.

Zigbee (2016). Disponível em < http://www.zigbee.org/> [Acessado em 28/06/2016].