Information Fusion 15 (2014) 90–101

Contents lists available at SciVerse ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

An approach to implement data fusion techniques in wireless sensor networks using genetic machine learning algorithms

A.R. Pinto [a,*], C. Montez [b], G. Araújo [b], F. Vasques [c], P. Portugal [d]

[a] UNESP, Universidade Estadual Paulista, SP, Brazil
[b] PGEAS, UFSC, Universidade Federal de Santa Catarina, SC, Brazil
[c] IDMEC, FEUP, Universidade do Porto, Porto, Portugal
[d] INESC TEC, FEUP, University of Porto, Porto Portugal

ARTICLE INFO

Article history:
Received 25 April 2010
Received in revised form 5 March 2013
Accepted 9 May 2013
Available online 18 May 2013

Keywords:
Partial data fusion
Autonomic computing
Classifier systems

ABSTRACT

Wireless Sensor Networks (WSNs) can be used to monitor hazardous and inaccessible areas. In these situations, the power supply (e.g. battery) of each node cannot be easily replaced. One solution to deal with the limited capacity of current power supplies is to deploy a large number of sensor nodes, since the lifetime and dependability of the network will increase through cooperation among nodes. Applications on WSN may also have other concerns, such as meeting temporal deadlines on message transmissions and maximizing the quality of information. Data fusion is a well-known technique that can be useful for the enhancement of data quality and for the maximization of WSN lifetime. In this paper, we propose an approach that allows the implementation of parallel data fusion techniques in IEEE 802.15.4 networks. One of the main advantages of the proposed approach is that it enables a trade-off between different user-defined metrics through the use of a genetic machine learning algorithm. Simulations and field experiments performed in different communication scenarios highlight significant improvements when compared with, for instance, the Gur Game approach or the implementation of conventional periodic communication techniques over IEEE 802.15.4 networks.

1. Introduction

Wireless Sensor Networks (WSNs) are composed of a large number of sensor nodes (also known as motes), usually with reduced size. Each node has a processor, memory, sensors, a wireless communication module and a power supply. In the most common situations, the power supply is provided by batteries. These networks can be deployed in hazardous areas, where maintenance is a difficult task. Therefore, energy consumption is a major concern because in these situations it is very difficult to replace batteries. Besides energy, applications for WSN may also have to deal with time-constrained messages [1].

The use of dense networks (i.e. with a high number of nodes per m^2), with hundreds or thousands of nodes, may be an advantage. A dense WSN can improve the accuracy of results, overpassing the unreliable nature of individual nodes. However, the usage of dense networks imposes more complex management approaches [1,2]. The use of autonomic approaches can be an interesting alternative for dense WSN, since self-management techniques [3] are adequate methodologies to manage large number of nodes.

In dense networks, the use of data fusion approaches may be an advantage due to the large number of messages exchanged, since a data fusion task may merge several messages in an useful and accurate information for the final user [4]. Additionally, some data fusion techniques proposed in the literature can deal with delays, message losses and discards, allowing greater flexibility in both the choice of the network technology and the communication approach used to disseminate data on the network [5].

Regarding the large number of nodes, the topology of network must be considered as being dynamic, as an unpredictable number of nodes can present hardware, software or communication failures. Thus, it is extremely difficult to obtain an accurate view of the network topology, which prevents the use of static scheduling of messages.

In WSN, applications may have multiple conflicting goals, such as timing constraints conflicting with energy consumption constraints. Another example may be the case of a data fusion application that requires a large number of messages to guarantee a minimum accuracy. Nevertheless, increasing the number of message transmissions can also increase the number of collisions. This leads both to an increase of the energy consumption due to message retransmissions and to a reduction of the number of messages that arrive to the final destination (e.g. message discards due to retransmission limits).

* Corresponding author.
E-mail addresses: arpinto@ibilce.unesp.br (A.R. Pinto), montez@das.ufsc.br (C. Montez).

1566-2535/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.inffus.2013.05.003

The IEEE 802.15.4 [6] has been used extensively as a communication infrastructure for WSN to support a large number of applications. These networks may be organized in peer-to-peer or star topologies. However, large coverage area and obstacles may prevent direct communication between nodes. Therefore, it may be impossible, in some circumstances, to implement star topologies. Nevertheless, star topologies constitute an interesting solution due to several reasons. First, star topologies are simpler to manage than other alternatives. Moreover, current RF transceivers compatible with IEEE 802.15.4 can operate in 868 MHz or 900 MHz bands reaching long distances in line-of-sight conditions, which makes feasible the establishment of large outdoor networks with direct communication with a base station. Finally, it is important to note that complex topologies based on cluster-tree approaches are ultimately formed by sets of clusters configured as star topologies.

In this paper, we present a generic approach that allows the implementation of parallel data fusion techniques in IEEE 802.15.4 networks based on a star topology. The main advantage of the proposal is the trade-off among different metrics through the use of a genetic machine learning algorithm – GMLA, which allows obtaining self-management properties in dense wireless sensor networks. The adopted communication scheme is based on local node decisions to define whether sensed information will be sent, or not, to the base station. The approach was developed considering requirements related with: dynamic topology changes, applications with conflicting goals and the impossibility of human intervention.

In order to assess the advantages of the proposed approach, a set of simulations was performed to compare it with both the Gur Game [7] and the use of a simple periodic application running over IEEE 802.15.4 standard [6]. Aiming to analyze the behavior and stability of the approach, different simulation scenarios were considered, varying the number of nodes (from few ones to thousands nodes), the deployment area (100×100 m and 1000×1000 m), and node mobility (nodes can enter or leave the radio range of the base station). Experiments with motes have also been performed with the purpose of to validate the simulation results in a real environment. They show that the envisaged approach is viable to be used on real scenarios.

The remainder of this paper is organized as follow. Section 2 presents the background related to this proposal. In Section 3 related works are discussed and the proposed model is introduced. The proposed approach is discussed in Section 4. In Section 5 the proposal is assessed and results are discussed. Finally, in Section 6 conclusions and final remarks are presented.

2. Background

2.1. IEEE 802.15.4

IEEE 802.15.4 [6] is becoming a *de facto* standard for low power and low rate WSNs. The physical layer can operate up to 250 Kbps of maximum transmission rate, working in different frequencies, such as the popular 2.4 GHz and 868/900 MHz bands for long range communication in Europe and North America, respectively. The MAC layer supports two operational modes that can be selected by the PAN coordinator: *beaconless mode*, using a non-slotted CSMA/CA mechanism; and *beacon mode*, where beacons are sent periodically by the PAN coordinator, and nodes are synchronized by a superframe structure. Since our proposal uses the beaconless mode, we restrict the following discussion to this operating mode.

Two variables are maintained by each device in beaconless mode: *NB*, the number of times CSMA/CA mechanism is required to backoff and *BE*, the backoff exponent, which is related to how many backoff periods a device shall wait before attempting to access a channel. There are also three main parameters: *macMinBE*, *macMaxBe* and *macMaxCSMABackoffs*, which corresponds, respectively, to the initial value of *BE*, the maximum value of *BE* and maximum number of backoffs before being declared a channel access failure.

The protocol works in the following manner. The *BE* is initially set to *macMinBE* value. When sending a message, the MAC sublayer shall delay for a random number of backoff periods in the range of $0–2 * BE - 1$, after which requests to the physical layer to perform a channel assessment. If the channel is assessed to be busy, the MAC sublayer will increment *NB* and *BE*. Once *BE* reaches to the *macMaxBE* value, it remains at this maximum value. If the value of NB is greater than *macMaxCSMABackoffs* (default value 4), the CSMA/CA shall end with a channel access failure status.

Therefore, *macMinBE*, *macMaxBe* and *macMaxCSMABackoffs* parameters can influence the beaconless CSMA/CA performance. These default values can decrease battery consumption (a device tries to retransmit only up to five times before aborting the transmission), however when the number of nodes increases, the communication efficiency drastically decreases [8,9]. Thus, the use of IEEE 802.15.4 protocol does not seem to be adequate for dense networks without addition of a mechanism such as the one proposed in this work.

2.2. Genetic algorithms and classifier systems

Classifier systems are machine-learning algorithms based on genetic algorithms [10]. These systems are able to learn simple rules, called classifiers. In this paper we call it Genetic Machine Learning Algorithm (GMLA), where "learning", in this context, means a *continuum* online adaptation process to a partially unknown and dynamic environment. Classifier systems are composed of three main components (Fig. 1): rules and message system; bank (usually called apportionment system); and genetic algorithm (GA).

The main idea behind a classifier system is to model the system, which represents the problem, by a set of condition-action rules. As a classifier system is a learning system based on Darwinian

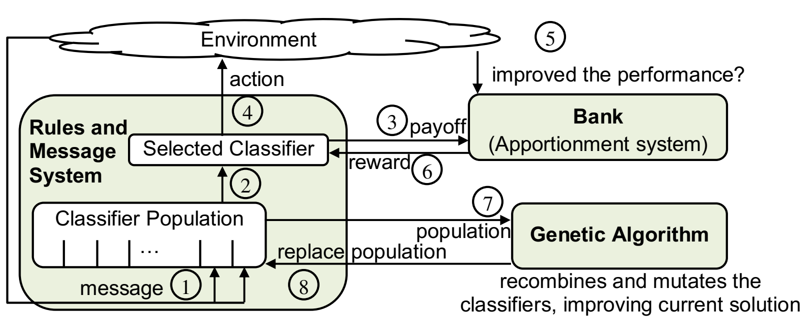

Fig. 1. Classifier system scheme.