Advanced Science and Technology Letters
Vol.62 (Sensor 2014), pp.33-36
http://dx.doi.org/10.14257/astl.2014.62.09

Efficient Construction of Decision Tree for Parallel Processing of Data from Wireless Sensor Networks

Aziz Nasridinov[1], Young-Ho Park[2*]

[1]School of Computer Engineering, Dongguk University at Gyeongju
123 Dongdaero, Gyeongju, Gyeongbuk, 780-714, Korea
aziz@dongguk.ac.kr
[2]Department of Multimedia Science, Sookmyung Women's University,
Cheongpa-ro 47 gil 100,Yongsan-gu, Seoul, 140-742, Korea
yhpark@sm.ac.kr
*Corresponding Author

Abstract. Potential worldwide deployment of WSNs for, e.g., environmental monitoring purposes could yield data in amounts of petabytes each year. Thus, in addition to the highly interesting technical challenges related to WSNs themselves, their widespread deployment would also require development of solutions for analyzing the potentially huge amounts of data they would generate. The contribution of this paper is two-fold. First, we propose an approach for constructing a decision tree based classification model for sensor data obtained from WSNs. Proposed model helps in analyzing uncovering patterns, associations, anomalies, and statistically significant structures and events in sensor data. Second, in order to speed up the performance of proposed decision tree based classification model, we propose ubiquitous parallel computing approach for construction decision tree on Graphic Processing Unit (GPU). Due to parallel features of GPU computing, we can accelerate the computing of decision tree algorithm.

Keywords: parallel computing, Graphic Processing Unit (GPU), decision tree.

1 Introduction

Wireless sensor networks (WSNs) are constructed of many tiny and low-cost sensor nodes randomly scattered over a large location. It can be used in many applications, such as military target tracking and surveillance, meteorological hazards, wildlife monitoring, and natural disaster relief. Most of the recent research in the field of WSNs has mainly focused on the technical challenges related to WSNs such as developing solutions for energy efficient deployment, routing and management of WSNs. However, potential worldwide deployment of WSNs for, e.g. environmental monitoring purposes could yield data in amounts of petabytes each year [3]. Thus, in addition to the highly interesting technical challenges related to WSNs themselves, their widespread deployment would also require development of solutions for analyzing the potentially huge amounts of data they would generate.

The contribution of this paper is two-fold. First, we propose an approach for constructing a decision tree based classification model for sensor data obtained from

ISSN: 2287-1233 ASTL

WSNs. Proposed model helps in analyzing uncovering patterns, associations, anomalies, and statistically significant structures and events in sensor data. Second, in order to speed up the performance of proposed decision tree based classification model, we propose ubiquitous parallel computing approach for construction decision tree on Graphic Processing Unit (GPU). Due to parallel features of GPU computing, we can accelerate the computing of decision tree algorithm.

The rest of the paper proceeds as follows. Section 2 discusses the related work. Section 3 describes the proposed method. Section 4 highlights conclusions.

2 Related Work

In this section, we describe related work towards the data analysis in WSNs. We first explain the GPU and its main components, and then describe the data processing in WSNs.

GPU is a ubiquitous device, which exists in every personal computing system. The modern GPU is not just a simple graphic manipulator but it gives it very high throughput on certain problems, and its near universal use in desktop computers means that it is a cheap and ubiquitous source of processing power [1]. There is increasing interest in applying this power to general-purpose problems through frameworks such as NVIDIA's Compute Unified Device Architecture (CUDA), an application programming tool developed to provide programmers a standard way to implement general-purpose applications on NVIDIA GPUs. We can exploit CUDA to accelerate computationally intensive data processing operations, often executing them fifty times faster on the GPU [7].

The authors in [3] explored the problems of storing and reasoning about data collected from very large-scale WSNs. In their paper, they provide a study on how suitable existing distributed database solutions, and especially the MapReduce programming model would be for the basis of such infrastructure for storing and analyzing sensor data. The authors [2] proposed a novel decision-tree-based hierarchical distributed classification approach, in which local classifiers are built by individual sensors and merged along the routing path forming a spanning tree. However, larger training sets enable the construction of more accurate decision trees, but also increase the processing and memory requirements of the tree building phase. These issues are not considered in aforementioned papers. Therefore, techniques that speed up decision tree building are needed.

3 Proposed Method

In this section, we describe the proposed method how to apply the proposed method in WSN and accelerate the performance of it.

Data mining techniques can be applied to facilitate the analyzing of data that is obtained from WSNs. One of the key steps in data mining is a classification task. Classification is the procedure to build a rule using the pre-defined classes and their features in dataset and apply this rule to a new data for discriminating each observation [5].

Decision tree is one of the well-known classification techniques. The ID3 algorithm [8] is a widely used data classification solution for decision tree learning. In a decision tree, each non-leaf node contains a splitting point, and the main task for building a decision tree is to identify the test attribute for each splitting point. The ID3 algorithm uses the entropy and information gain to select the test attribute. These calculations are performed on GPU that can leave the space for other calculations for CPU. Information gain can be computed using entropy. The entropy is calculated using the following formula [6]:

$$entropy(D) = -\sum_{j=1}^{|c|} \Pr(c_j) log_2 \Pr(c_j) \qquad (1)$$

where $Pr(c_j)$ is the probability of class c_j in data set D. If we make attribute A_i, with v values, the root of current tree, this will partition D into v subsets $D_1, D_2, \ldots, D_v$. which is performed on GPU device. Information gained by selecting attribute A_i to branch or to partition the data is as following [6], where we choose the attribute with the highest gain to branch/split the current tree:

$$gain(D, A_i) = entropy(D) - entropy_{A_i}(D) \qquad (2)$$

In this paper, we applied the proposed method mentioned above in the WSNs field in order to process the data obtained from it. Traditionally, the data obtained from the WSNs are collected in the base station, and then base station filters out the information and sent it to the application. However, large amounts of data collected every second can make the processing time-wasting and can reduce the energy of base station. The proposed method can facilitate the work of base station by constructing the decision tree for the data obtained from the WSNs. For example, if WSN monitors the temperature of a particular region, the proposed method can classify the temperature of regions by hot, rainy, cloudy, etc. It can be useful not only in knowledge discovery, that is, the identification of new phenomena, but also it can help in enhancing our understanding of known phenomena. On the other hand, one of the main advantages of the proposed method is that we propose a GPU based acceleration for our method. Due to parallel features of GPU computing, we can accelerate the computing of decision tree algorithm.

We also propose a performance measurement method of the proposed method. We have carried out the performance measurement that compares the proposed method with the well-known machine learning classification methods, such as support vector machine (SVM), k-nearest neighbor (kNN), etc. Particularly, we have compared the proposed method with those algorithms implemented on solely CPU, implemented solely on GPU, and implemented on hybrid CPU-GPU approaches. The result of experiments show that hybrid CPU-GPU approach outperforms CPU-based sequential implementation by up to four orders times. Thus, the hybrid CPU-GPU method not only accelerates the construction of decision tree via GPU computing, but also does so in the context of characterizing the power and energy consumption of the GPU. Due to the shortage of the space, we will describe the experiment results in detail in full paper.

4 Conclusion

In this paper, we first proposed an approach for constructing a decision tree based classification model for sensor data obtained from WSNs. Second, in order to speed up the performance of proposed decision tree based classification model, we proposed ubiquitous parallel computing approach for construction of decision tree on GPU. In the future, we plan to perform a performance evaluation of the proposed method. Particularly, we plan to measure the accuracy and overall processing time of the proposed method compared to the well-known machine learning classification algorithms, including SVM, kNN, etc.

Acknowledgement. This work was supported by the IT R&D program of MKE/KEIT. [10041854, Development of a smart home service platform with real-time danger prediction and prevention for safety residential environments].

References

1. Bakum, P. and Skadron, K. (2010), ‘Accelerating SQL Database Operations on a GPU with CUDA’ In GPGPU’10: Proceedings of the Third Workshop on General-Purpose Computation on Graphics Processing Units, Pittsburgh, PA, USA, pp. 94-103.
2. Cheng, X., Xu, J., Pei, J., and Liu, J. (2010), ‘Hierarchical Distributed Data Classification in Wireless Sensor Networks’ Computer Communication, Vol. 33 No. 15, pp. 1404-1413.
3. Jardak, C., Riihijärvi, J., Oldewurtel, F. and Mähönen, P. (2010), ‘Parallel processing of data from very large-scale wireless sensor networks’ in HPDC '10: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, Chicago, Illinois, USA, pp. 787-794.
4. Larose, D.T. (2005), Discovering knowledge in data, John Wiley & Sons.
5. Lim, N. (2007), Classification by Ensembles from Random Partitions using Logistic Regression Models. Stony Brook University, New York, USA.
6. Nasridinov, A., Park, Y. H. (2014), Decision Tree Construction on GPU: Ubiquitous Parallel Computing Approach, Computing, Vol. 96, No. 5, pp. 403-413, 2014.
7. Shuai, C., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J.W. and Skadron, K. (2008), ‘A performance study of general-purpose applications on graphics processors using CUDA’ Journal of Parallel and Distributed Computing, Vol. 68 No. 10, pp. 1370-1380.
8. Quinlan J. R. (1986), ‘Induction of Decision Trees’ Machine Learning, Vol.1 No.1, pp. 81-106.