Energy Consumption in Data Analysis for On-board and Distributed Applications

Ruchita Bhargava RUCHITA1@CS.UMBC.EDU
Hillol Kargupta HILLOL@CS.UMBC.EDU
Michael Powers MPOWER3@CS.UMBC.EDU
Department of Computer Science and Electrical Engineering University of Maryland Baltimore County, MD 21250

Abstract

Energy consumption is an important issue in the growing number of data mining and machine learning applications for battery-powered embedded and mobile devices. It plays a critical role in determining the capabilities of a broad range of applications such as space probes with on-board scientific missions, PDA-based monitoring of remote data streams, event detection in sensor networks comprised of battery-powered data sensors and light-weight data processing nodes. This paper presents an experimental investigation of the energy consumption characteristics of different data analysis techniques. The paper compares the energy consumption characteristics of common data analysis operations on-board a mobile device with the energy necessary to send the same amount of data over wireless networks to a remote machine for analysis. It benchmarks the performance and points out that the energy consumption for transmitting data over low bandwidth lossy wireless channels often supersedes that for the operations on-board a light-weight compute node.

1. Introduction

Many autonomous space applications deal with a voluminous stream of scientific data. Analyzing this data is often critical for the underlying scientific mission. Since transmitting large amount of data to the remote control station for subsequent analysis is usually very expensive and often not practical for time-critical applications, modern space applications are increasingly relying on autonomous on-board data analysis. Examples of such type of application environments include sensor networks/webs (Srivastava et al., 2001), on-board satellite-based platforms (Chien et al., 2001), on-board vehicle monitoring system, mobile health monitoring system (Jorge, 2001). These applications often face many resource constraints. In addition to the demand of low bandwidth usage and the challenges of limited computing resources, minimizing energy consumption is another important requirement in these devices.

This paper focuses on the energy efficiency of data mining algorithms in applications developed to run on small embedded and possibly mobile devices. Most of these autonomous data collection/processing devices are battery operated and lack a constant source of power. For example in sensor networks the battery life could be equal to the life of the device as some of these devices may not be recharged. Most commercially available mobile computing devices like PDA's and mobile phones have battery power which would last for only a few hours. For on-board satellite applications, the power which is generated using solar cells and stored in batteries for later use has to be utilized judiciously. Therefore, the next generation of data mining applications for such embedded and mobile devices must be designed to minimize the energy consumption. Software power utilization and minimization has been studied in various contexts (Roy & Johnson, 1996; Tiwari et al., 1996; Flinn & Satyanarayanan, 1999) but to the best of our knowledge there does not exist a study on energy requirements for data mining algorithms.

Based on the experimental energy consumption data collected for several popular data analysis techniques, this paper introduces energy efficiency as a novel criterion to evaluate algorithms intended to run on mobile and embedded platforms. This paper reports three sets of experiments:

- Characterizing the energy consumption behavior for the traditional centralized architecture of transmitting the data to a remote site for mining.
- Identifying the power consumption characteristics of some commonly used statistical and data mining algorithms running on-board a mobile device.
- Characterizing the energy consumption behavior of a

distributed data mining algorithm.

We characterize and compare the energy consumed by data mining applications implemented on-board a device with the energy costs for sending the data to a remote site. This paper shows that for low bandwidth lossy networks the high energy costs of communication often makes local on-board data mining a more energy efficient choice. Although the results are intuitive, this paper documents the first effort to quantify the power consumption characteristics of monolithic and distributed data mining algorithms. In doing so, it motivates the need for designing power efficient on-board data analysis applications.

The rest of the paper is organized as follows. Section 2 introduces the importance of power conservation in implementation of data mining algorithms on portable and embedded devices. Section 3 discusses the experimental setup used for the energy measurements. Section 4 discusses the results obtained for various data mining algorithms. Section 5 presents the conclusions and discusses the future work.

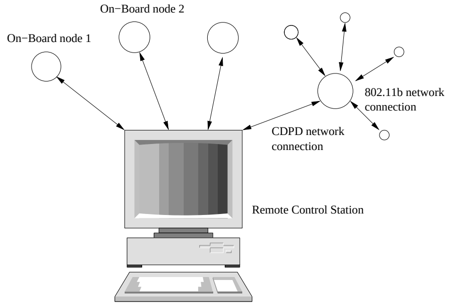

Figure 1. Experimental network configuration that this paper studies.

2. Energy consumption in Mobile Devices

Most embedded devices and sensors do not have a constant renewable source of power. The power consumed by these devices during the execution of a program is defined as the rate at which energy is consumed. Although the terms power and energy are sometimes used interchangeably in the literature, usually we are concerned with energy efficiency as batteries have a finite supply of energy. The power consumed by a process can be defined as $P(t) = V(t) \times I_c$, where $V(t)$ is the supply voltage, I_c is the current, and t stands for the time index. On the other hand, energy consumption is defined as $E = \int_{t_1}^{t_2} P(t)dt$.

Therefore, a technique that reduces power consumption will save energy only if it does not increase the execution time by a factor that exceeds the gains from the power reduction. The total energy consumption of a system depends on both the hardware and the software components. Energy utilization characteristics of a system depend on the computation and the communication load. For a data mining application implemented on an embedded mobile device we define the total energy utilized to be $E = E_c + E_t$, where E is the total energy needed, E_c is the energy needed for computation and E_t is the energy needed for communication of data.

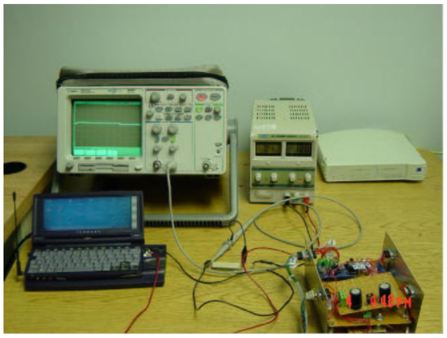

Figure 2. Experimental setup for measuring energy consumption.

This paper explores the power consumption characteristics of a collection of distributed and centralized data mining algorithms. The application environment studied in this paper is illustrated in Figure 1. It consists of a collection of remote data processing and data collection nodes that are connected to a remote control station through low-bandwidth wireless networks. The remote nodes (represented by circles in the figure) represent the on-board data processing nodes. Some of the nodes may themselves be connected through relatively faster wireless networks in a peer-to-peer mode.

The DIADIC laboratory at UMBC [1] is currently working on several distributed applications that fit the scenario outlined in Figure 1. For example, we are working on an on-board mobile data mining application that analyzes a continuous stream of real time data collected from a vehicle to monitor its health and performance. In this application the on-board data processing and collection nodes are mobile. We are also working on a distributed data mining application for a grid computing environment (Hingne et al., 2003) that involves “light-weight” sensor nodes. We have also developed the MobiMine system (Kargupta et al., 2002) for remotely mining data streams from handheld devices for mobile users.

Most off-the-shelf data mining systems work in a centralized fashion. If the data is distributed then it must be downloaded to a central location before it can analyzed. This constraint poses a critical challenge in mining distributed data sources. Is the current technology appropriate for these new breed of distributed wireless applications?

The following section explores this question from the per-

[1] Distributed Adaptive Discovery and Computation Group at UMBC (http://www.cs.umbc.edu/ hillol/diadic.html)

spective of energy consumption. It studies the energy consumption characteristics of several popular centralized and distributed data analysis techniques.

3. Experimental Setup

This section describes the experimental set-up used for the energy consumption measurements. The data set used for our experiments was collected as a part of the experimental setup for a PDA based mobile vehicle health monitoring system. Data was collected for sixty-four real-valued features with every observation consisting of 256 bytes of data. Amongst others, these features include engine control data, vehicle diagnostic data and emission control data. The energy measurement experiments are performed using (1) a HP Jornada 690 (Hitachi SuperH SH-3 processor with 32MB RAM running Windows CE 2.11 build 9018), Compaq 802.11b card, Sierra Wireless CDPD Aircard 300 card, and a Pentium IV (2GHz with Windows 2000) as the control station. Figure 2 shows the actual experimental setup.

An embedded Visual C++ compiler was used to compile the code with all compiler optimizations turned off. The battery from the unit was removed and a 12V DC power was applied. Avoiding the batteries allows for a more steady voltage to be supplied to the device. Energy consumption was determined by measuring the input voltage and current across a test resistance of 1 Ω using a Agilent 54622A oscilloscope. The reported values were averaged over five independent runs using the same data set. We also measured the base energy requirement of the HP Jornada without any application executing on it. This was the energy needed to run the WinCE tasks and the graphical user display. We subtracted these base values from our experimental results to get the true values.

4. Results

This section, presents the experimental results for energy consumption characteristics of the data mining algorithms implemented on our experimental testbed.

4.1. Communication and Energy Consumption

The choice of communication network has a substantial effect on energy consumption. We have chosen two different networks for our computations and they are briefly explained in the following. Wireless LAN (IEEE 802.11b) is a communication technology operating at 2.4 GHz frequency with a data transfer rate upto 11Mbps. A Compaq WL110 Wireless PC Card was used for these experiments. This card draws a current of 185mA in receive mode and 285mA in transmit mode.

Cellular Digital Packet Data (CDPD) is a specification for supporting data transmission on cellular phone frequencies. CDPD transmits data at rates of up to 19.2 Kbps using cellular channels in the range of 800 - 900 MHz. Sierra Wireless Aircard 350 CDPD modem with typical transmit current draw of 500 mA was used for our experiments.

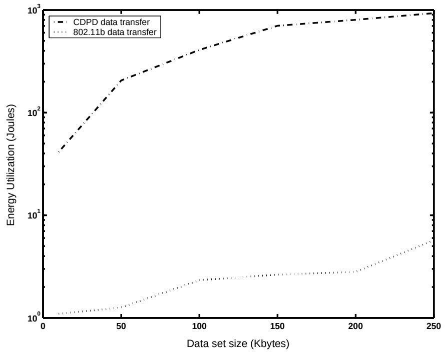

Figure 3. Energy consumption in communicating data to a remote site without any local computation using CDPD and 802.11b ethernet LAN. The average standard deviation for 802.11b and CDPD is 0.1364 and 71.67 respectively.

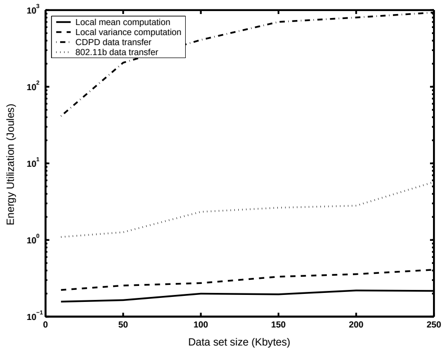

Figure 4. Comparison of energy utilized to communicate data to remote site with no local computation and energy used in calculating mean and variance locally. The average standard deviation for mean and variance computation is 0.0877 and 0.1281 respectively.

The power consumption for the data transmission process for a mobile node transmitting the data to a remote site from the UMBC location using these two technologies is shown in Figure 3. As we expect, Figure 3 shows that the amount of energy needed for communicating data to a remote site using a lossy low bandwidth CDPD network is considerably higher compared to that needed using a 802.11b network. The lossy nature of the CDPD network combined with the fact that the time needed to transmit data depends on the availability of the network causes the variance of the measured values for CDPD network to be high.

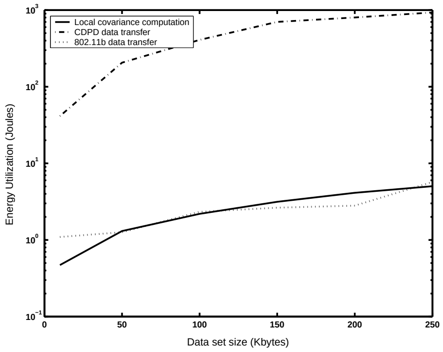

Figure 5. Comparison of energy utilized to communicate data to remote site with no local computation and energy used in calculating covariance matrix locally. The average standard deviation for covariance computation is 0.0176.

4.2. Energy Consumption for Data Mining

4.2.1. Basic Statistical Aggregates

Most data mining applications require computing some basic statistical aggregates such as mean, variance, and covariance matrix. Therefore, we should first explore the energy consumption behavior of the algorithms to compute them. The energy consumed in computing these aggregates locally at the HP Jornada using our experimental setup are shown in Figure 4 and 5. The figures also show the energy needed to transmit the same amount of data over the CDPD and the 802.11b network from a remote on-board node to the central control station. The results show that the the energy needed to compute these aggregates on-board is considerably lower than the energy needed for communicating the data over a lossy low bandwidth CDPD channel. The covariance matrix is relatively compute-intensive and its on-board computation appears to be comparable to the performance of the procedure that first transmits the data over the 802.11b network.

4.2.2. Principal Component Analysis (PCA)

PCA (Hotelling, 1933) is a popular technique to construct a representation of the data set that reduces its dimensionality by transforming the set of correlated variables into a set of transformed uncorrelated variables called principal components. Usually we choose the top few eigenvectors capturing the maximally variant dimensions to project the data into a new reduced dimensional space. For our experiments we use the Householder reduction method to calculate the eigenvectors and eigenvalues of the covariance matrix. The power utilization for performing PCA on-board the HP Jornada is profiled in Figure 6. It shows that the energy consumption is least when the data is directly transmitted to the processing node over 802.11b connection. The performance of the on-board computation scheme appears to be better than CDPD but worse than 802.11b.

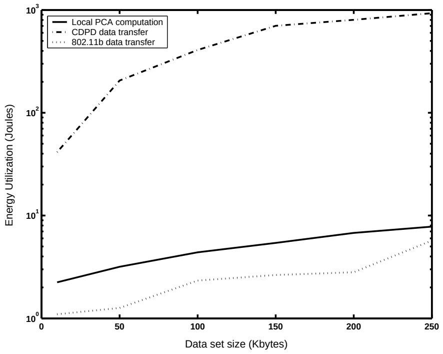

Figure 6. Comparison of energy utilized to communicate data to remote site with no local computation and energy used in calculating PCA locally. The average standard deviation for PCA computation is 0.4128.

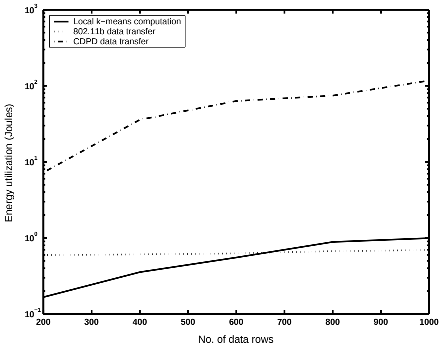

Figure 7. Comparison of energy utilized to communicate data to remote site with no local computation and energy used in clustering data using K-means locally. The average standard deviation for k-means clustering computation is 0.0265.

4.2.3. K-Means Clustering

Clustering involves division of a data set into smaller subsets consisting of similar objects. The k-means algorithm (Hartigan & Wong, 1979) is a popular clustering tool used for machine learning applications. K-means creates k clusters of the data where every cluster can be represented by the mean of its data points, called the centroid. For our experiments we have used a data set consisting of seven features collected from the experimental mobile on-board vehicle system. We use an iterative k-means clustering method which assigns all the points to their nearest centroids followed by recomputation of the centroids until a stopping criterion is met. In Figure 7 we show the energy needed for k-means clustering computation on-board the HP Jornada when k is set to three. When implemented on-board, this simple clustering algorithm is a much more energy efficient choice than communicating the data over the CDPD network.

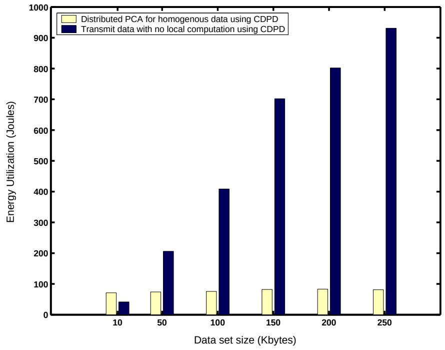

Figure 8. Comparison of energy consumption at distributed nodes for Centralized and Distributed PCA for homogeneous data using CDPD network. The average standard deviation for distributed PCA computation is 6.6408.

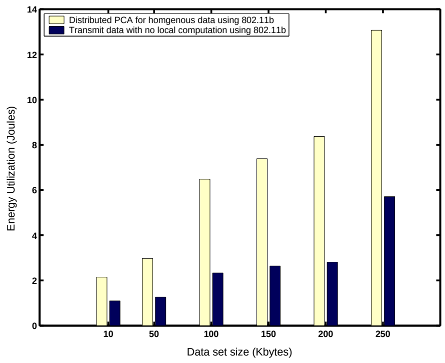

Figure 9. Comparison of energy consumption at distributed nodes for Centralized and Distributed PCA for homogeneous data using the 802.11b network. The average standard deviation for distributed PCA computation is 0.3038.

4.2.4. Distributed and Centralized PCA

This section explores a distributed data mining algorithm (Park & Kargupta, 2002) from the energy consumption perspective. It particularly considers the problem of performing PCA from distributed homogeneous and heterogeneous data. The most obvious solution is to download the data to a central location and then perform the PCA. The distributed techniques try to accomplish this without downloading all the data. The distributed algorithms differ depending on the type of the distributed data sets—homogeneous or heterogeneous.

Homogeneous data sites share the same set of features. In this case, the central control station can perform PCA if it has knowledge of the covariance matrices for all the nodes. We can compute the covariance of the local data sets at the distributed nodes and transmit the covariance matrix to the central site. Transmission of the covariance matrix would suffice for global PCA computation at a central site since covariance is additive. Figure 8 and Figure 9 show the energy comparisons for centralized and distributed PCA for homogeneous data for CDPD and 802.11b networks respectively. It clearly shows that the energy consumption performance of the distributed algorithm is significantly better than downloading the data to the central station for CDPD but for a high bandwidth network like 802.11b, the computation intensive PCA algorithm is an expensive choice in terms of energy.

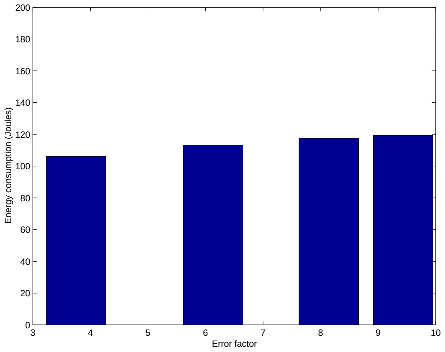

Figure 10. Energy consumption for varying error values (corresponding to varying sampling rates of local projected data) for heterogeneous distributed PCA. The average standard deviation for distributed Heterogeneous PCA computation is 7.3993.

Now let us consider the case where the distributed data sites are heterogeneous (Park & Kargupta, 2002). In other words, each site contains data sets defined over different subsets of features linked by one or more overlapping key features. For example, consider a scenario where a company manufactures two different models of the car but uses the same engine. The task is to find how the engine performs in the two different models. In this case the data is collected from two different models and hence does not consist of the same features. For such heterogeneous data, we use the Collective PCA approach outlined in (Kargupta et al., 2000) to calculate PCA at the central site. Local PCA is performed at each site and the dominant eigenvectors are chosen to project the data. For n total data rows a sample of $c << n$ projected data rows along with the eigenvectors of the covariance matrix are transmitted to a central site. The fact that PCA is invariant to linear transformations, is exploited to perform global PCA at the central site on the combined projected data instead of the original data. The collective PCA technique introduces approximations in the computation of the global covariance matrix by transmitting only a subset of projected data to the central site. For our experiments we ignore the error introduced due to the dimension reduction of the data at the distributed sites since it is small in the reported experiments. The collective PCA technique exploits local PCA computation and subsequent sampling of the local data in the projected space for sending a relatively small amount of data to the central site. More details can be found elsewhere (Kargupta et al., 2000). The error introduced in the global covariance matrix $C_{x'}$, when m rows are chosen for sampling as compared to the true covariance matrix C_x is given by $||C_{x'} - C_x||_F$, where $||.||_F$ denotes the Frobenius norm.

As we increase m, the energy cost of transmitting the data increases. For our experiments, we create two subsets of the data with each one stored at a different distributed node. Following the Collective PCA approach, we perform local PCA at each site and set the principal component selection threshold to 98%. Global PCA is performed at the central site for different sampling rates and the energy consumption profile for different error rates (corresponding to different sampling rate) is presented in Figure 10. Both the distributed algorithms perform much better than their centralized counterpart.

5. Conclusions and Future Work

Energy efficiency is an important design consideration in development of data mining algorithms for mobile and distributed environments. In this paper, we have tried to experimentally quantify the performance of specific data mining algorithms from the energy consumption perspective. The paper first points out that for most common data analysis techniques on-board computation is a better option when the nodes are connected through low bandwidth wireless networks like CDPD. However, for nodes connected through the 802.11b LAN network, transmitting the data to a remote data processing node may be an equally good option when the data analysis operations are relatively compute-intensive. The paper also points out that traditional centralized data mining techniques may not be appropriate for distributed applications where the nodes are connected through low bandwidth networks like CDPD. It compared the performance of the centralized PCA with that of distributed PCA for both homogeneous and heterogeneous distributed data nodes using algorithms reported in the literature. The results show that the distributed algorithms work a lot better than their centralized counterpart from the power consumption perspective.

We are currently exploring the energy consumption characteristics of various other data mining algorithms. We are also decomposing the algorithmic performance in terms of computing primitives and exploring the possibility of identifying higher level algorithmic design principles that may be useful for developing the next generation of power efficient on-board distributed applications.

6. Acknowledgments

The authors acknowledge support from the United States National Science Foundation CAREER award IIS-0093353, NASA (NRA) NAS2-37143, and TEDCO, Maryland Technology Development Center.

References

Chien, S., Engelhardt, B., Knight, R., Rabideau, G., Sherwood, R., Hansen, E., Ortiviz, A., Wilklow, C., & Wichman, S. (2001). Onboard autonomy on the three corner sat mission. *Proceedings of the 2001 International Symposium on Artificial Intelligence, Robotics, and Automation for Space*. Montreal, Canada.

Flinn, J., & Satyanarayanan, M. (1999). Energy-aware adaptation for mobile applications. *Proceedings of the Symposium on Operating Systems Principles* (pp. 48–63).

Hartigan, J., & Wong, M. (1979). A k-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Hingne, V., Joshi, A., Finin, T., Kargupta, H., & Houstis, E. (2003). Towards a pervasive grid. *Proceedings of the Next Generation Systems Program Workshop, Next Generation Systems Program Workshop*. Nice, France.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*.

Jorge, A. (2001). Adaptive tools for the elderly: new devices to cope with age-induced cognitive disabilities. *Proceedings of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing: providing for the elderly* (pp. 66–70).

Kargupta, H., Huang, W.Krishnamurthy, S., & Johnson, E. (2000). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems Journal*, *3*, 422–448.

Kargupta, H., Park, B., Pittie, S., Liu, L., Kushraj, D., & Sarkar, K. (2002). Mobimine: Monitoring the stock market from a PDA. *ACM SIGKDD Explorations*, *3*, 37–46.

Park, B. H., & Kargupta, H. (2002). Distributed data mining: Algorithms, systems, and applications. In N. Ye (Ed.), *The handbook of data mining*. Lawrence Erlbaum Associates.

Roy, K., & Johnson, M. (1996). Software design for low power. *NATO Advanced Study Institute on Low Power Design in Deep Submicron Electronics*.

Srivastava, M., Muntz, R., & Potkonjak, M. (2001). Smart kindergarten: sensor-based wireless networks for smart developmental problem-solving enviroments. *Proceedings of the seventh annual international conference on Mobile computing and networking* (pp. 132–138).

Tiwari, V., Malik, S., Wolfe, A., & Lee, M.-C. (1996). Instruction level power analysis and optimization of software. *Journal of VLSI Signal Processing Systems*, *13*.