Supporting a Real-Time Distributed Intrusion Detection Application on GATES

Qian Zhu, Liang Chen, and Gagan Agrawal

Department of Computer Science and Engineering
The Ohio State University, Columbus OH 43210
{zhuq, chenlia, agrawal}@cse.ohio-state.edu

Abstract. Increasingly, a number of applications across computer sciences and other science and engineering disciplines rely on, or can potentially benefit from, analysis and monitoring of data streams. We view the problem of flexible and adaptive processing of distributed data streams as a grid computing problem. In our recent work, we have been developing a middleware, GATES (Grid-based Adaptive Execution on Streams), for enabling grid-based processing of distributed data streams.

This paper reports an application study using the GATES middleware system. We focus on the problem of intrusion detection. We have created a distributed and self-adaptive real-time implementation of the algorithm proposed by Eskin using our middleware. The main observations from our experiments are as follows. First, our distributed implementation can achieve detection rates which are very close to the detection rate by a centralized algorithm. Second, our implementation is able to effectively adjust the adaptation parameters.

1 Introduction

Increasingly, a number of applications across computer sciences and other science and engineering disciplines rely on, or can potentially benefit from, analysis and monitoring of data streams. In the stream model of processing, data arrives continuously and needs to be processed in real-time, i.e., the processing rate must match the arrival rate. There are several trends contributing to the emergence of this model. First, scientific simulations and increasing numbers of high precision data collection instruments (e.g. sensors attached to satellites and medical imaging modalities) are generating data continuously, and at a high rate. The second is the rapid improvements in the technologies for Wide Area Networking (WAN). As a result, often the data can be transmitted faster than it can be stored or accessed from disks within a cluster.

The important characteristics that apply across a number of stream-based applications are: 1) the data arrives continuously, 24 hours a day and 7 days a week, 2) the volume of data is enormous, typically tens or hundreds of gigabytes a day, and the desired analysis could also require large computations, 3) often, this data arrives at a distributed set of locations, and all data cannot be communicated to a single site, 4) it is often not feasible to store all data for processing at a later time, thereby, requiring analysis in real-time.
We view the problem of flexible and adaptive processing of distributed data streams as a grid computing problem. We believe that a distributed and networked collection of computing resources can be used for analysis or processing of these data streams. Computing resources close to the source of a data stream can be used for initial processing of the data stream, thereby reducing the volume of data that needs to be communicated. Other computing resources can be used for more expensive and/or centralized processing of data from all sources.

In our recent work, we have been developing a middleware for enabling grid-based processing of distributed data streams [6,5]. Our system is referred to as GATES (Grid-based AdapTive Execution on Streams). One of the important characteristic of this middleware is that it can enable an application to achieve the best accuracy, while maintaining the real-time constraint. For this, the middleware allows the application developers to expose one or more adaptation parameters. An adaptation parameter is a tunable parameter whose value can be modified to increase the processing rate, and in most cases, reduce the accuracy of the processing. Examples of such adaptation parameters are, rate of sampling, i.e., what fraction of data-items are actually processed, and size of summary structure at an intermediate stage, which means how much information is retained after a processing stage. The middleware automatically adjusts the values of these parameters to meet the real-time constraint on processing, through a self-adaptation algorithm. Self-adaptation algorithms currently implemented in the middleware are described in our earlier papers [6,5].

This paper reports an application study using the GATES middleware system. We focus on the problem of intrusion detection, which a widely studied problem in computer security and data mining [1]. We have created a distributed and self-adaptive real-time implementation of the algorithm proposed by Eskin [3]. This implementation generates local models using data received at each node, and then combines these local models to create a global model. We use the functionality of GATES in two different ways. First, as network records typically arrive at multiple locations, a flexible distributed implementation can avoid high communication costs associated with a centralized implementation. Second, as data arrival rates can vary significantly, it is important for an intrusion detection implementation to choose the right trade-off between accuracy and processing rate, to continue to meet real-time constraints.

We have carried-out a number of experiments to evaluate our distributed implementation. The main observations from our experiments are as follows. First, our distributed implementation can achieve detection rates which are very close to the detection rate by a centralized algorithm. Second, our implementation is able to adjust the adaptation parameters. When the rate of data arrival is low, it chooses a small value of the adaptation parameter, EM convergence threshold, resulting in the best detection rate. On the other hand, when the data arrival rate is very high, it chooses a larger value of this parameter, resulting in somewhat lower accuracy, but still maintaining the same rate of processing.