Augmenting OSDC's Datascope with a Finderscope

Joseph Korpela
University of California Los Angeles

Matt Greenway
Laboratory for Advanced Computing

ABSTRACT

A key issue that has surfaced in a number of different contexts for the OSDC datascope is the need for the average user of this tool to be able to gain better insight into what data is available to them and into what correlations and patterns may exist in that data. Our proposal addresses this issue by providing all users with a prebuilt image which incorporates the RapidMiner toolset along with the needed support infrastructure that provides users with the ability to rapidly perform data preprocessing and analysis through an easily learned graphical user interface.

1. INTRODUCTION

While the users of OSDC may be well educated users with much experience in their own particular field, it is likely that many will still lack the experience needed to perform a wide ranging exploration of the overall dataset, if the only tools available to them are command line driven tools such as R. A tool based on a graphical user interface (GUI), such as RapidMiner, provides these users with a sort of finder scope, through which they can quickly get views of different parts of OSDC's dataset and perform initial analysis of that data, which will in turn allow them to better integrate new data sets into their research. In fact, the interface on such a tool is so well designed, that even experienced programmers will benefit from the ability to quickly try different combinations of preprocessing steps, filters, and machine learning algorithms on the various data sets.

2. OUR PROPOSAL

In more concrete language, our proposal is to build an image that will be available to all users, which is loaded with RapidMiner and all needed support infrastructure. This is not as simple as just installing RapidMiner to the instance, as running a GUI from a remote virtual machine (VM) brings with it a few challenges.

The foremost of these challenges is the need to forward the X11 terminal data (i.e. the graphics that make up the GUI) from the VM to the user's client computer. At first glance, this appears to be a simple issue of tunneling this data through the SSH session to the user's computer, but unfortunately the protocol used by X11 was not designed for use over long distances, and is extremely susceptible to lag. In our initial testing, we found that the lag between user input and remote response to be well over 10 seconds. The most likely solution to this challenge is to run the VM using virtual network computing (VNC). This system is designed specifically to allow remote access to a graphical user desktop, and as such, will allow users to interface with a GUI remotely with minimal lag.

While this appears to be a solution to our problem, it is not actually an easy one to implement. The OSDC as it stands does not allow for users to interact with a VNC server running on a VM directly, as the users are actually connecting first to a head node before accessing the actual VM. So, in order to utilize a VNC, it would either be necessary to open at least one VM up to an external IP address to allow remote access, or to install a VNC server on the head node, for the users to interface with the VNC at that level.

3. CONCLUSION

While other submissions may advocate for similar ideas, what we are proposing is for a specific implementation that meets a real need and which will increase the ease of use of OSDC, leading to increased exploration and discovery of OSDC's data. RapidMiner is a well known toolset that provides a host of well tested machine learning algorithms, combined with data preprocessing tools that allow users to take raw data and perform functions such as filtering, discretization, correlation analysis, and clustering, and does this through an easily learned interface. This could be implemented as a sort of introduction image, which both introduces users to the data and quickly gets them started in their discoveries.