ADAMS

Advanced **D**ata mining **A**nd **M**achine learning **S**ystem

Module: adams-weka-hadoop
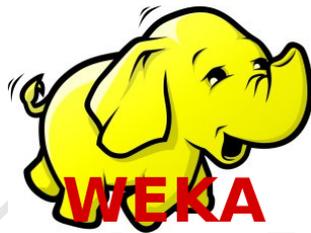
Zufeng Yu
Peter Reutemann

March 30, 2012

Contents

List of Figures

Chapter 1

Introduction

This manual includes all details regarding to run Hadoop experiments on both command line and Hadoop Experimenter from ADAM. It contains two major sections which are **Running From Command Line** and **Running from Hadoop Gui Experimenter**. There are two more sections afterwards which are **Tips on Hadoop Cluster setting** and **Summary**.

Chapter 2

Running From Command Line

In order to run weka experiments using hadoop from command line, there are 11 options have to be specified. The complete command line should look liks this:

```
bin/hadoop
Must be under hadoop directory

--config
Path of Hadoop configuration folder, etc /home/z123/hadoop-0.20.2/conf.

jar
The jar file generated by Hadoop Gui Experimenter on the fly. For example,
jar hadoopGui4812493.jar.

-libjars
All jars on classpath. For example, -libjars a.jar,b.jar,c.jar

-dataset
The path of an input dataset, can be used multiple times to input datasets.
For example, -datasets /home/datasets/a.arff -datasets /home/datasets/b.arff

-classifier
The path of an input classifier, can be used multiple times to input
classifiers. For example, -classifier weka.classifiers.functions.SMO
-classifier weka.classifiers.classifiers.trees.J48

-runs
Number of repetition, etc -runs 10

-folds
Number of folds, etc -folds 10

-exptype
Choice of {classification,regression}. Etc -exptype classification

-classindex
Choice of {last,first,default,an integer}. For example, -classindex last, or
-classindex 5, or -classindex default

-confhome
Path of the hadoop conf folder. The input of this option must be exactly the
same as --config.

-csv
Experiment output file path. Note that an arff file with same path and name
will be generated at same time. For example, by setting -csv /home/temp.csv,
you will get temp.csv and temp.arff under /home.
```

Every command option must be filled with correct input. When you run experiment on ADAMS using Hadoop experiment, it will generate a complete command line string each time it starts the hadoop experiment, we strongly suggest you to copy the full command line string instead of writing your own.

The jar file that you need is generated on the fly. You have to start the experiment on ADAMS using Hadoop Experiment, and the jar file will be created under hadoop home directory you have chosen. Note that there is no need to complete the experiment, because the program will create a jar file before everything else starts running, you can abort the experiment as soon as you get the jar file. The name and path of jar file will be shown on the panel.

Once you have the jar file, you can start running experiment from command line. The first three command options have to stay the exact order as shown above, which are `hadoop --config ...jar ...-libjars ....` And rest of the options are not restricted in order, but it is compulsory to provide input values to all the options.

Chapter 3

Running From Hadoop Gui Experimenter

After you start up Hadoop Experiment from ADAM, you will see a Gui interface similar with normal Weka experimenter. However there are a few changes need to pay attention to.

There are only two tabbed pannel instead of 3. We have removed the result pannel.

In **Setup** tab, under **Path Setting** section, it allows you to choose the hadoop home directory, specific hadoop configuration folder and the output file path. You can choose different versions of hadoop home directory, and each hadoop configuration may represent different cluster settings. Note that current experiment will rely on the configuration setting you have chosen. Regarding to output file path, it only asks you to give a path for CSV file, and the program will generate an arff file with same name/path in the end.

Hadoop Experimenter only performs Cross-Validation experiments, and loops iteration control will always be datasets first.

In **Run** tab, after you click on Start button, the experiment won't start until the program successfully create a jar file for current experiment.

Chapter 4

Tips on Hadoop Cluster

This chapter describes a few problems that might occur while running Hadoop Experiment on cluster. For more information please read Hadoop: The Definitive Guide [5], or Pro Hadoop [6].

A. Abort hadoop experiment

If a runnin hadoop experiment is interrupted half-way, it is necessary to run this command in command window: `hadoop --config .../conf job -kill currentJobId`. This command will kill all ongoing hadoop child process in the specified cluster for current Job. Job Id was shown on the panel when experiment started.

B. Java heap size error

By default hadoop only gives around 200m heap size to each task. This error might occur if you are running regression algorithms, such as SMOreg. The best solution so far is to increase Java heap size for each task. In the configuration folder, modify `conf/mapred-site.xml` file with few more lines:

```
<property>
    <name>mapred.child.java.opts</name>
    <value>-Xmx512m</value>
  </property>
```

It sets heap size to 512m, feel free to increase the size if necessary.

C. java.net.SocketTimeoutException: 480000 millis timeout

This bug sometimes occurs while running large experiment. It seems to be an I/O issue, and you can see the message in the tasktracker log files. However the only possible solution so far is to add following lines into conf/hdfs-site.xml file:

```
<property>
    <name>dfs.datanode.socket.write.timeout</name>
    <value>0</value>
</property>
```

D. Start up different clusters using `--config .../conf`

Normally you can use `bin/start-dfs.sh` and `bin/start-mapred.sh` to start
up a hadoop cluster, if you add `--config .../conf` after these commands, it
will start the specific hadoop cluster according to the conf setting. For example,
`bin/start-dfs.sh --config clusterA/conf`, `bin/start-mapred.sh --config`
`clusterA/conf`,now you have clusterA running, and then you can do `bin/start-dfs.sh`
`--config clusterB/conf`, `bin/start-mapred.sh --config clusterB/conf`,
and now you have clusterB running as well. Use `hadoop --config .../conf`
`dfsadmin -report` to check if you have the HDFS running. Also you can check
log files in log folder to see if everything works fine. There are a few types
of log files, and the files worth checking are related to namenodes, datanode,
jobtracker and tasktracker.

E. log file errors

Sometimes hadoop experimenter reports error in log files. One way to fix it
is to find out on which node has the problem. Then log in to that machine, under
hadoop home directory, if it's tasktracker' problem thenrun `bin/hadoop-daemon.sh`
`--config .../conf start tasktracker`. It will start tasktracker on this indi-
vidual machine, then it becomes part of the cluster based on your configuration
folder setting. If it is the datanode problem, then do `bin/hadoop-daemon.sh`
`--config .../conf start datanode`.

F. Better view of hadoop process

Suppose the master machine is ml64-20.cms.waikato.ac.nz, and the port for
jobtracker is 50030. Then log in to ml64-20, start a browser, and type in
http://ml64-20.cms.waikato.ac.nz:50030. You will have an overall view of the
cluster, about how many nodes available, what is current job status etc.

Chapter 5

Summary

It is strongly recomended that using Hadoop Gui Experimenter from ADAMS to run experiments. The program has been designed to automatically remove all the unnecessary files that were generated during a hadoop process, except for the final output files and the jar file.

Multiple experiments can be run simutaneously on a cluster, or on several clusters. As long as you provide different output file names to those experiments, it shouldn't be a problem.

Bibliography

[1] *ADAMS* – Advanced Data mining and Machine learning System
`http://adams.cms.waikato.ac.nz/` `http://adams.cms.waikato.ac.nz/`

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.
`http://www.cs.waikato.ac.nz/ml/weka/`

[3] Ian H. Witten, Eibe Frank, Mark A. Hall (2011); *Data Mining: Practical Machine Learning Tools and Techniques*; Third Edition; Morgan Kaufmann; ISBN 978-0-12-374856-0
`http://www.cs.waikato.ac.nz/ml/weka/book.html`

[4] *Apache Hadoop* – Open-source software for reliable, scalable, distributed computing
`http://hadoop.apache.org/`

[5] Tom White (2009); *Hadoop: The Definitive Guide*; First Edition;O'Reilly; ISBN 978-0-596-52197-4
`http://books.google.com/books?id=Nff49D7vnJcC`

[6] Jason Venner(2009); *Pro Hadoop*; APress; ISBN 978-1-4302-1942-2
`http://books.google.com/books?id=H3mvcxPeUfwC`