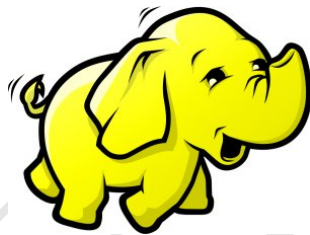


ADAMS

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

Module: adams-weka-hadoop



Zufeng Yu
Peter Reutemann

January 27, 2012

©2012



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Contents

1	Introduction	7
2	Running From Command Line	9
3	Running From Hadoop Gui Experimenter	11
4	Tips on Hadoop Cluster	13
5	Summary	15
	Bibliography	17

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 string should look like this:

(under hadoop home directory) bin/hadoop

-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 Tll 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.

In order to run Hadoop experiment successfully, 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 GUI.

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

Chapter 4

Tips on Hadoop Cluster

Java heap size

Can run multiple experiments on same cluster, but it would be better to run on different cluster.

Bug: `java.net.SocketTimeoutException: 480000 millis timeout set`

Chapter 5

Summary

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/>