Cloud Computing Homework 2

Sirui Xing

Submitted Nov 27th, 2014

Problem description

The purpose of this assignment is to gain experience using Mahout and HBase. You are given two data sets: daily weather and birth statistics for the U.S. in 2012. It is up to you to come up with one hypothesis to test and subsequently determine whether that hypothesis stands.

Data Set overview

The daily weather and birth data sets are limited to the U.S. for the period starting January 1st, 2012 to December 31st, 2012. These data have been placed into HBase on shark. A copy is also available in hdfs at /cosc6376/homework2/ for reference purposes only. It is expected that you will be working with HBase directly for at least your first MapReduce task.

The four tables for each relevant data set are accessible through HBase:cosc6376-hw2-weather, cosc6376-hw2-birth, cosc6376-hw2-lunar and cosc6376-hw2-stations.

Please note: For your first MapReduce job you are expected to query these tables.

Procedure Descriptions

It may be too complicated to cross analysis data in different tables. I decide to focus on weather table, and particularly the precipitation data set. I would like to study the precipitation pattern that observed at different stations. Hoping that the precipitation pattern can be classified into a few individual groups. The data vector would be all year round precipitation value of each station. To reduce array dimension for better clustering, I use the mean precipitation of each month, which is 12 data in total. The clustering program I use is mahout kmeans. Before clustering, an optimal cluster number is calculate using Matlab eva = evalclusters(x,clust,criterion,Name,Value)

It turns out that for my data, cluster number k=3 is the optimal. The distance measurement is Euclidean. The larger Euclidean distance means the larger absolute precipitation difference. After mahout kmeans, I study means and standard deviation of each cluster. The final step is to do a hypothesis test. Let's denote the mean of global January precipitation is μ_0. The claim being investigated is that the average January precipitation of a particular cluster group is greater than μ_0.

Step 1. Reading data from HBase

The purpose of the first step is simply to get data from HBase. Using TableMapper.class from HBase Java API, a mapreduce job is able to read the desired data from HBase and write into HDFS. The source code is *ReadHBase.java*. Mapper will read HBase line by line. Each line contains information of station id, date, highest temperature value, lowest temperature value, precipitation value, and some other flags. The only data we need are station id, date, and precipitation value. The

output is a list of key value pairs. The key is station id, and the value is a string of date and precipitation in year 2012.

$ *hbase shell scan 'cosc6376-hw2-weather'*

```
AE000041196-20120124-PRCP    column=cf:station_id, timestamp=1415488981747, value=AE000041196
AE000041196-20120124-PRCP    column=cf:type, timestamp=1415488981747, value=PRCP
AE000041196-20120124-PRCP    column=cf:value, timestamp=1415488981747, value=0
AE000041196-20120124-TMAX    column=cf:date, timestamp=1415488981747, value=20120124
AE000041196-20120124-TMAX    column=cf:m-flag, timestamp=1415488981747, value=
AE000041196-20120124-TMAX    column=cf:obs_time, timestamp=1415488981747, value=
AE000041196-20120124-TMAX    column=cf:q-flag, timestamp=1415488981747, value=
AE000041196-20120124-TMAX    column=cf:s-flag, timestamp=1415488981747, value=S
AE000041196-20120124-TMAX    column=cf:station_id, timestamp=1415488981747, value=AE000041196
AE000041196-20120124-TMAX    column=cf:type, timestamp=1415488981747, value=TMAX
AE000041196-20120124-TMAX    column=cf:value, timestamp=1415488981747, value=207
AE000041196-20120124-TMIN    column=cf:date, timestamp=1415488981747, value=20120124
AE000041196-20120124-TMIN    column=cf:m-flag, timestamp=1415488981747, value=
AE000041196-20120124-TMIN    column=cf:obs_time, timestamp=1415488981747, value=
AE000041196-20120124-TMIN    column=cf:q-flag, timestamp=1415488981747, value=
AE000041196-20120124-TMIN    column=cf:s-flag, timestamp=1415488981747, value=S
AE000041196-20120124-TMIN    column=cf:station_id, timestamp=1415488981747, value=AE000041196
AE000041196-20120124-TMIN    column=cf:type, timestamp=1415488981747, value=TMIN
AE000041196-20120124-TMIN    column=cf:value, timestamp=1415488981747, value=100
AE000041196-20120125-PRCP    column=cf:date, timestamp=1415488981747, value=20120125
```

Part of HBase table 'cosc6376-hw2-weather'

Step 2. Convert text file to sequence file.

In order to run Mahout, the input vector has to be sequence file of particular format. There are two utilities come along with mahout intended for format conversion. One is *seqdirectory*, the other is *seq2sparse*. After some study of these two utilities, I realized they have more suitable for search indexing, and not intended for general application. So I wrote my own conversion code. The source code is *CreateSequenceFile.java*. The output is *prcp_namedvector_month.seq*. The key is station id, and the value includes (1) station id, (2) a list of precipitation data with its position. Each precipitation data is the mean of the month.

```
Key: AG000060390: Value:
AG000060390:{0:9.0,1:82.79310344827586,2:26.0,3:60.793103448275865,4:7.967741935483871,5:0.6,7:12.87096774193
5484,8:8.5,9:27.129032258064516,10:31.482758620689655,11:15.161290322580646}
Key: AG000060590: Value:
AG000060590:{0:9.033333333333333,2:1.4137931034482758,3:1.0714285714285714,9:2.966666666666667,10:0.333333
3333333333}
Key: AG000060611: Value:
AG000060611:{1:0.6896551724137931,2:1.6451612903225807,10:0.16666666666666666,11:0.6451612903225806}
Key: AG000060680: Value:
AG000060680:{5:0.3333333333333333,6:0.25806451612903225,7:1.1666666666666667,8:1.5,9:0.6451612903225806,10:
0.3333333333333333,11:1.0}
```

Part of converted sequence file

Step 3. Kmeans cluster

Finding optimal number of cluster using Matlab evalclusters function.

eva = evalclusters(prcp_data,'kmeans','CalinskiHarabasz','KList',[1:20])

eva =

CalinskiHarabaszEvaluation with properties:

NumObservations: 14748
InspectedK: [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20]
CriterionValues: [1x20 double]
OptimalK: 3

Mahout kmeans

mahout kmeans -i ./input/ -o ./output -c init_centers -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure -cd 0.01 -k 3 -x 100 -ow –cl

- cluster number is set to be 3.
- convergence delta is 0.01.
- distance measure is Euclidean

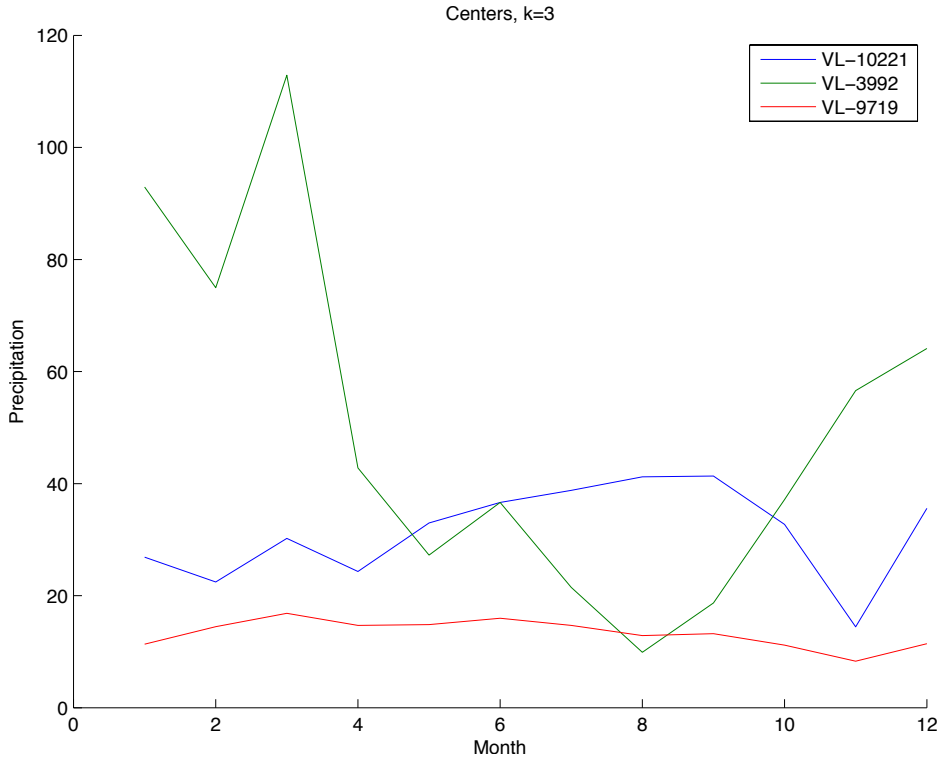
clusterdump

mahout clusterdump -i ./output/clusters-5-final/ --pointsDir ./output/init_centers -o ./cluster-test-1-0.01.out

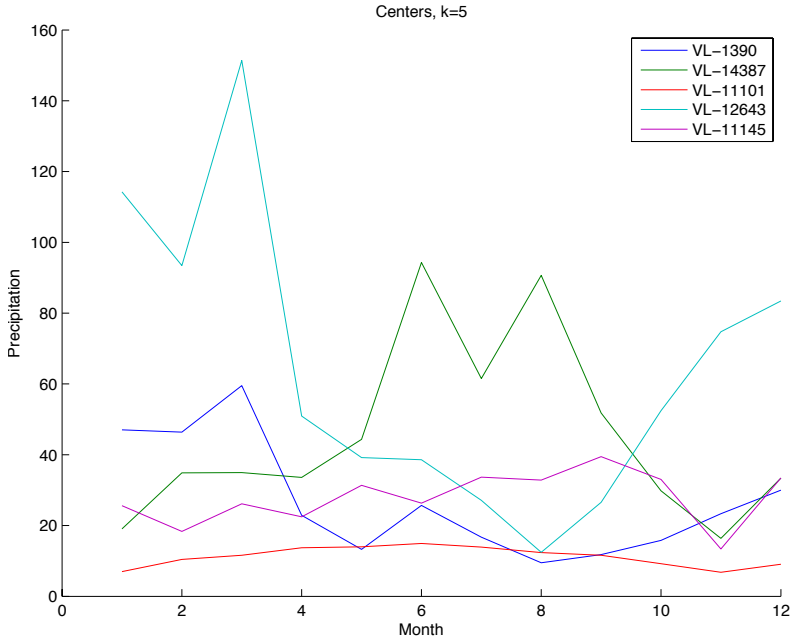
The clusterdump utility is to convert sequence result into text file. For this data set, the cluster result is:

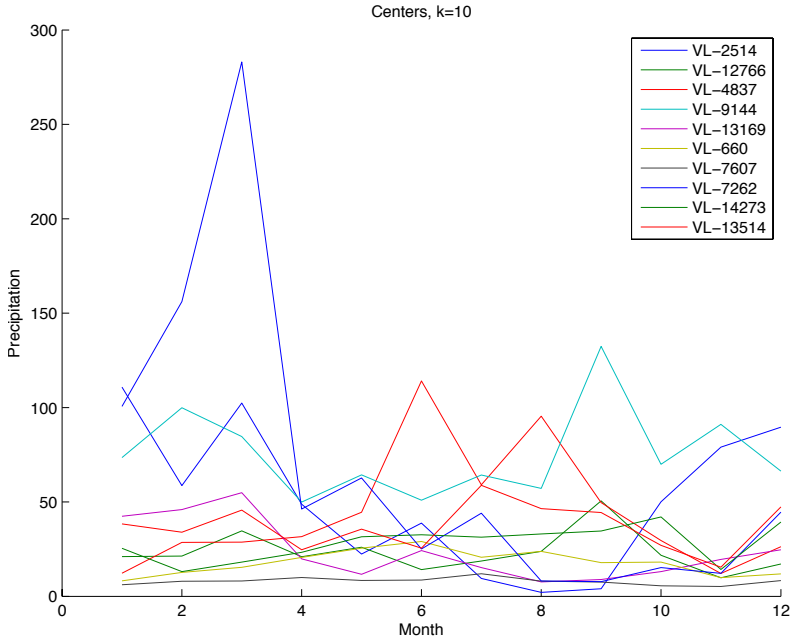
VL-10221{n=5259 c=[26.858, 22.447, 30.213, 24.302, 32.977, 36.625, 38.789, 41.207, 41.338, 32.708, 14.436, 35.598] r=[15.200, 16.682, 19.604, 14.320, 17.883, 33.319, 23.558, 27.921, 22.695, 18.202, 13.528, 17.244]}
VL-3992{n=920 c=[92.942, 74.962, 112.886, 42.798, 27.229, 36.637, 21.472, 9.893, 18.687, 37.136, 56.587, 64.117] r=[50.487, 44.571, 73.666, 26.835, 29.791, 30.568, 25.065, 20.927, 42.377, 35.732, 43.247, 44.503]}
VL-9719{n=8569 c=[11.337, 14.472, 16.842, 14.700, 14.823, 15.975, 14.690, 12.888, 13.205, 11.165, 8.299, 11.436] r=[14.408, 15.185, 17.790, 12.294, 13.829, 13.658, 12.016, 11.325, 13.082, 10.773, 8.766, 11.482]}

The three clusters center ids are VL-10221, VL-3992, and VL-9719 respectively. The VL-10221 has n=5250 data belonged to. The center is c=[26.858, 22.447, …35.598]. The centers are plot using Matlab.



As comparison, I also did k=5, and k=10. These clustering results are showed below.





From these plot, we can have an idea that there is one class that has really high precipitation around spring and winter (e.g. VL-3992). Some classes have a relatively high precipitation during summer (e.g. 10221). The rest of the classes have the same precipitation all year round (e.g. 9719).

Step 4. Hypothesis Test:

VL-3992 group has an average January precipitation that is higher than the overall average January precipitation of μ_0.

A random statistical sample of 100 January precipitation data from VL-3992 group is selected. The average January precipitation of these data is found to be μ_1, with standard deviation of σ_1.

The claim being investigated is that the average January precipitation of VL-3392 group is greater than μ_0. This corresponds to the statement $x \geq \mu_0$.

The null hypothesis H_0: $x = \mu_0$. Alternative hypothesis H_1: $x > \mu_0$

μ_0: 21.9621

σ_0 27.3949 (generally this value is not available, and we have to use σ_1 to evaluate)

μ_1: 104.8639

σ_1: 41.0527

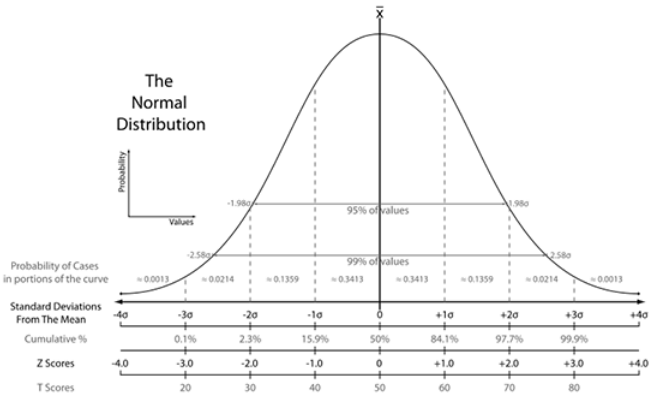
The 100 samples are large enough that we can safely assume they form Gaussian distribution. The z value can be formulated by:

$$Z = \frac{\mu 1 - \mu 0}{\sigma 1/\sqrt{100}}$$

$$Z = 82.9018/4.1053 = 20.193847$$

Suppose that a = 0.05. We can draw the appropriate picture and find the z score for -0.025 and 0.025. We call the outside regions the rejection regions.

Since z is so high, the probability that H_0 is true is so small that we decide to reject H_0 and accept H_1.



Z value larger than 3 means it falls outside the 3σ region.

As a conclusion, the alternative hypothesis is accepted, H_1: $x > \mu_0$. VL-3992 group has an average January precipitation that is higher than the overall average January precipitation of μ_0.