University of CINCINNATI

BIG DATA ANALYTICS

CS 7070

Homework – 1 (MR Decision Tree Example)

Submitted on:

03/14/2019

Submitted by

Siva Sai Krishna Paladugu

Problem Statement:

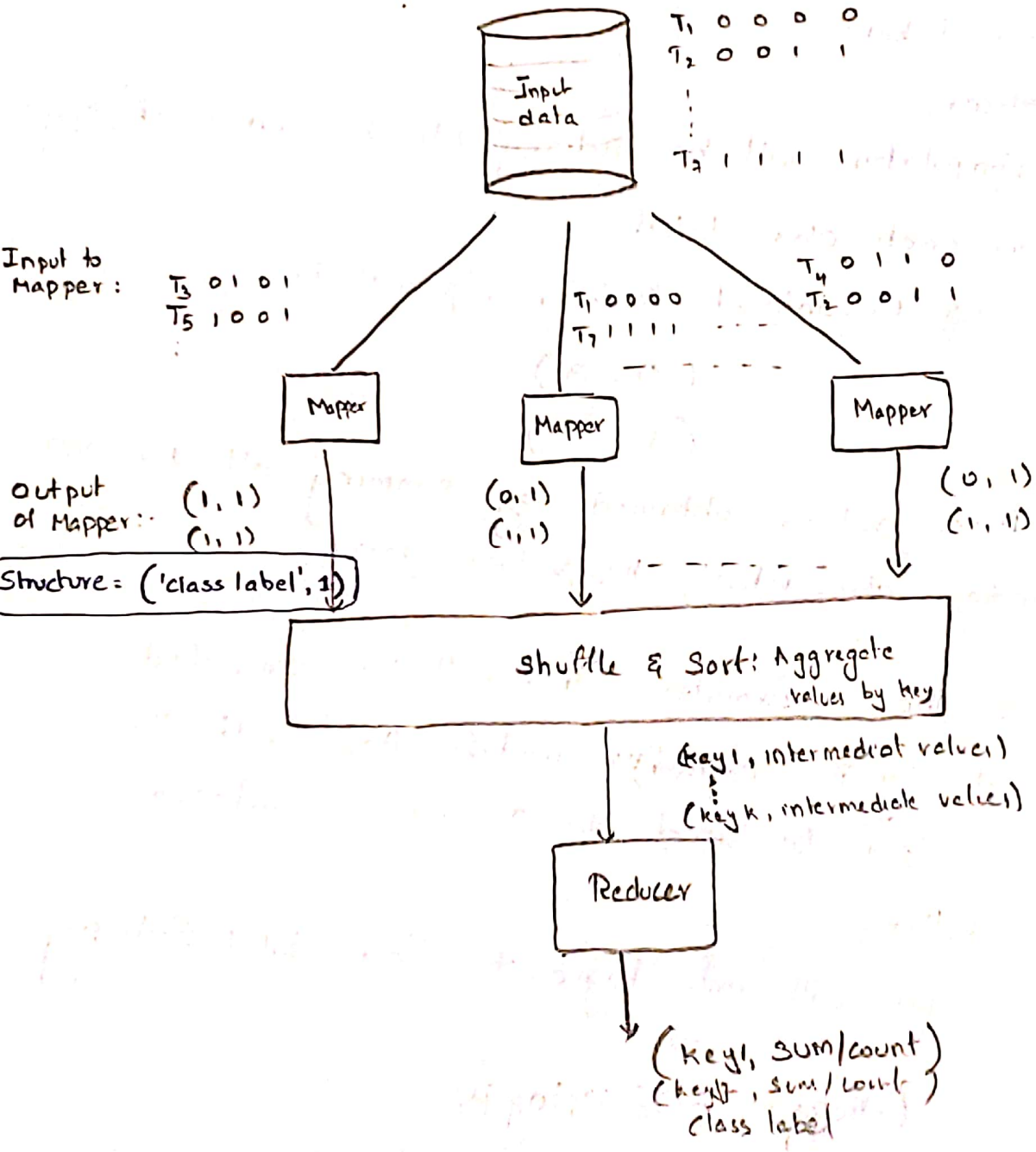
Consider the data shown in the table. This data is to be used to construct a decision tree. We discussed in class today the outline of a MapReduce algorithm that will build a decision tree from a very large dataset stored in in HDFS across multiple nodes. We want to use the ID3 algorithm for decision tree induction that uses information gain to select the best attribute at test node of a decision tree. Assume that there is a controller program that wants to build the decision tree by launching various MapReduce jobs, and using and saving requisite results after each MapReduce iteration. In this context answer the following questions:

Dataset Table

	A1	A2	A3	Class
T1	0	0	0	0
T2	0	0	1	1
T3	0	1	0	1
T4	0	1	1	0
T5	1	0	0	1
T6	1	1	0	0
T7	1	1	1	1

1. The controller launches a MapReduce iteration to compute the basic entropy of this database.

a) Describe the structure of key-value pairs to be generated by the Mapper.



b) Describe the computation performed by the Reducer.

Ans: Here,

From mapper we will get two unique key and their values (for given dataset)

- keys are class labels i.e., 0 & 1
- values corresponding to keys will be 1's & their number is equal to no. of records with corresponding class labels.

In Reducer,

Computation will be adding all 1's for each unique key

i.e., for each class label.

eg: for dataset, Reducer o/p will be

(0, 3)

(1, 4)

→ These values obtained by summing all 1's from corresponding class label's key, values pairs.

---

c) Describe the information that will be computed and saved by the controller module. How will the reducer output be used to do this computation.

Ans: In controller,

we will write logic to find total Entropy

$$\text{Entropy} = -\sum P_i \log_2 P_i$$

If two class labels,

$$E = -(P_1 \log_2 P_1 + P_0 \log_2 P_0)$$

$$= -\left(\frac{n_1}{n} \log_2 \frac{n_1}{n} + \frac{n_0}{n} \log_2 \frac{n_0}{n}\right)$$

Where,
$n \rightarrow$ Total no. of records.
$n_0 \rightarrow$ no. of records with class label-0
$n_1 \rightarrow$ no. of records with class label-1

The same logic can be applied if multiple classes present by taking more terms like $n_0, n_1, \ldots$

For given data set,

$$n_1 = 4 \quad , \quad n_0 = 3$$

$$\therefore E = -\left(\frac{4}{4+3} \log_2 \frac{4}{4+3} + \frac{3}{4+3} \log_2 \frac{3}{4+3}\right)$$

$$= 0.46134 + 0.52388 = \boxed{0.9852}$$

This entropy will be saved by the controller for future use.

Reducer o/p $\Rightarrow$ (0, 3), (1, 4) is used by controller to get n_0 & n_1 and to calculate total entropy.

Note :- If multiple reducers are present, then in controller we will write logic to combine all outputs from all reducers depending on key. (same case if multiple iterations used)

d. Show all the key-values pairs generated by Mapper for the shown dataset?

Ans: Given records,

	A_1	A_2	A_3	class
T_1	0	0	0	0
T_2	0	0	1	1
T_3	0	1	0	1
T_4	0	1	1	0
T_5	1	0	0	1
T_6	1	1	0	0
T_7	1	1	1	1

Mapper output:- ('class label', 1) → structure

(0, 1)
(1, 1)
(1, 1)
(0, 1)
(1, 1)
(0, 1)
(1, 1)

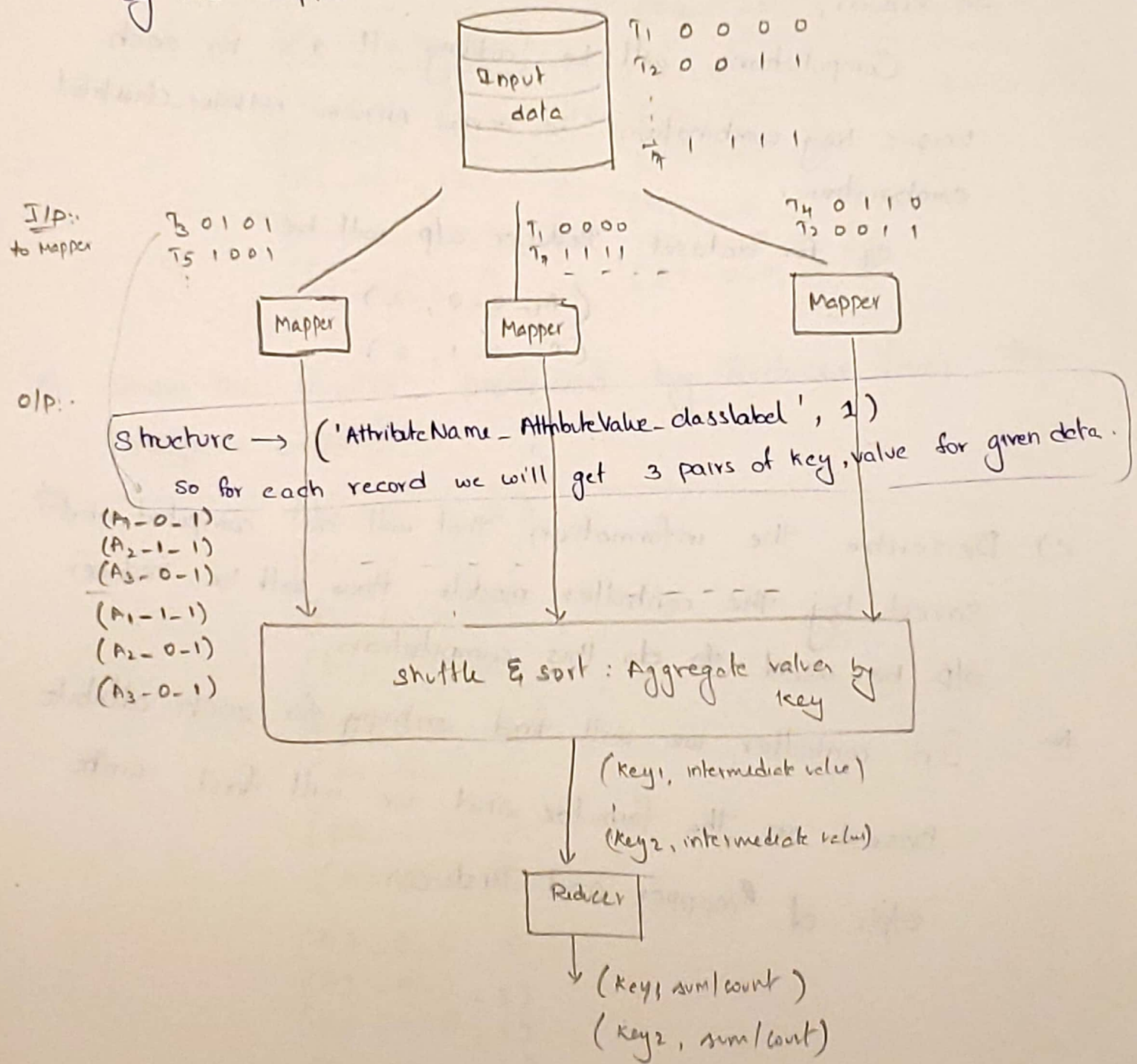
e. Show the results produced by Reducer from the Mapper's output?

Ans:- Reducer o/p:-

(0, 3)
(1, 4)

2. The controller launches a MapReduce iteration to determine the best test attribute, from among the three attributes of the dataset. We want to achieve this with only one iteration of MapReduce.

a) Describe the structure of key-value pairs to be generated by the Mapper.



b. Describe the compilation performed by the Reducer

Ans: Here,

From mapper we will get combination of Attribute names, Attribute values and class labels.

Keys: AttrName - AttrValue - Class label

Value: 1

In Reducer;

Computation will be adding all 1's for each unique key combination. ie, unique AttName_AttValue_Class label combination.

eg:- for dataset, Reducer o/p will be

(A1-0-0, 2)

(A1-0-1, 2)

⋮

c.) Describe the information that will be computed and saved by the controller module. How will be reducer o/p be used to do this computation.

Ans: In controller, we will find entropy for each attribute based on the formulae, and we will first write o/p's of mapper and Reducer

d. Show all the key-values pairs generated by mapper for the shown dataset?

Ans: Mapper o/p:- List of key value pairs.
structure → (AttrName - AttrValue - classlabel , 1)

T_1
(A1 - 0 - 0, 1)
(A2 - 0 - 0, 1)
(A3 - 0 - 0, 1)

T_2
(A1 - 0 - 1, 1)
(A2 - 0 - 1, 1)
(A3 - 1 - 1, 1)

T_3
(A1 - 0 - 1, 1)
(A2 - 1 - 1, 1)
(A3 - 0 - 1, 1)

T_4
(A1 - 0 - 0, 1)
(A2 - 1 - 0, 1)
(A3 - 1 - 0, 1)

T_5
(A1 - 1 - 1, 1)
(A2 - 0 - 1, 1)
(A3 - 0 - 1, 1)

T_6
(A1 - 1 - 0, 1)
(A2 - 1 - 0, 1)
(A3 - 0 - 0, 1)

T_7
(A1 - 1 - 1, 1)
(A2 - 1 - 1, 1)
(A2 - 1 - 1, 1)

e. Show the results produced by Reducer from the Mappers output?

Ans.

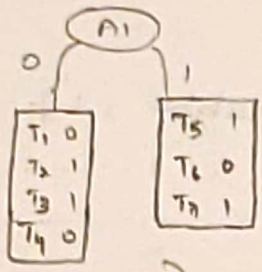
(A1 - 0 - 0, 2)
(A1 - 0 - 1, 2)
(A1 - 1 - 0, 1)
(A1 - 1 - 1, 2)

(A2 - 0 - 0, 1)
(A2 - 0 - 1, 2)
(A2 - 1 - 0, 2)
(A2 - 1 - 1, 2)

(A3 - 0 - 0, 2)
(A3 - 0 - 1, 2)
(A3 - 1 - 0, 1)
(A3 - 1 - 1, 2)

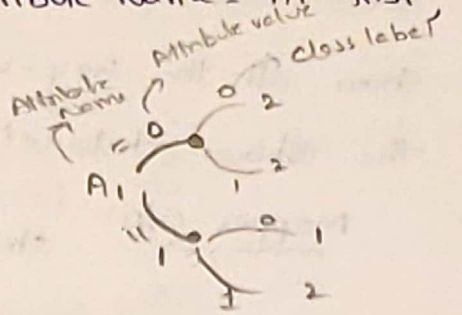
We will split the keys from the outputs of Reducer by '-' and consider all key-value pairs with Attribute Name = A1 first,

So,



(for reference)

(A1_0_0 , 2)
(A1_0_1 , 2)
(A1_1_0 , 1)
(A1_1_1 , 2)



$$E(A1=0) = -\frac{n_0^0}{n^0}\log_2\frac{n_0^0}{n^0} - \frac{n_1^0}{n^0}\log_2\frac{n_1^0}{n^0}$$

Where $n^0 \rightarrow$ Total records with $A1=0 = n_0^0 + n_1^0$

$n_0^0 \rightarrow$ no. of records with $A1=0$ & class $= 0$

$n_1^0 \rightarrow$ no. of records with $A1=0$ & class $=1$

$$= -\frac{2}{2+2}\log_2\frac{2}{2+2} - \frac{2}{2+2}\log_2\frac{2}{2+2}$$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= -1\log_2 2^{-1} = \boxed{1}$$

$$E(A_1=1) = -\frac{n_0'}{n'}\log_2\frac{n_0'}{n'} - \frac{n_1'}{n'}\log_2\frac{n_1'}{n'}$$

where

$n' \rightarrow$ Total records with $A1=1 \Rightarrow n_0' + n_1'$

$n_0' \rightarrow$ no. of records with $A1=0$ & class $=0$

$n_1' \rightarrow$ no. of records with $A1=1$ & class $=1$

$$= -\frac{1}{1+2}\log_2\frac{1}{1+2} - \frac{2}{1+2}\log_2\frac{2}{1+2}$$

$$= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.5283 + 0.3899 = 0.91827$$

Average Entropy Information for A1,

$$I(A1) = \frac{N_0'}{N'} * E(A1=0) + \frac{N_1'}{N'} * E(A1=1)$$

Where
N_0' → No. of records with A1 = 0
N_1' → No. of records with A1 = 1
N' → Total no. of records.

$$I(A1) = \frac{4}{7}(1) + \frac{3}{7}(0.91827)$$

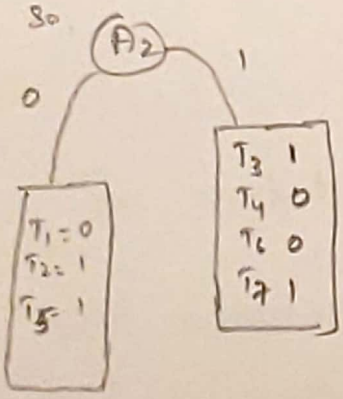
$$= 0.96497$$

$$Gain(A1) = E - I(A1)$$ (Here E value taken from ques 1)

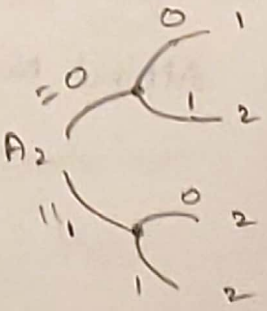
$$= 0.9852 - 0.96497$$

$$\boxed{Gain(A1) \simeq 0.0202} \rightarrow (1)$$

Now will split and get key-value pairs with Attribute Name = A_2



(A2 - 0 - 0, 1)
(A2 - 0 - 1, 2)
(A2 - 1 - 0, 2)
(A2 - 1 - 1, 2)



$$E(A2=0) = -\frac{1}{1+2}\log_2\frac{1}{1+2} - \frac{2}{1+2}\log_2\frac{2}{1+2}$$ (using same formula but now considering A2)

$$= 0.5283 + 0.3899$$

$$= 0.91827$$

$$E(A2=1) = -\frac{2}{2+2}\log_2\frac{2}{2+2} - \frac{2}{2+2}\log_2\frac{2}{2+2}$$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= 1$$

Now,

$$I(A2) = \frac{3}{7}(0.91827) + \frac{4}{7}(1)$$ (Now considering A2 in records)

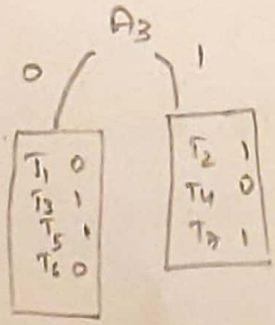
$$= 0.96497$$

$$Gain(A2) = E - I(A2)$$

$$= 0.9852 - 0.96497$$

$$\boxed{Gain(A2) \simeq 0.0202} \longrightarrow (2)$$

Now we will split the keys and key-value pairs with Attribut Name = A3 are considered,



(A3_0_0, 2)
(A3_0_1, 2)
(A3_1_0, 1)
(A3_1_1, 2)

$$E(A3=0) = \frac{-2}{2+2}\log_2\frac{2}{2+2} - \frac{2}{2+2}\log_2\frac{2}{2+2}$$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= 1$$

$$E(A3=1) = -\frac{1}{1+2}\log\frac{1}{1+2} - \frac{2}{1+2}\log\frac{2}{1+2}$$

$$= 0.5283 + 0.3899$$

$$= 0.91827$$

$$I(A3) = \frac{4}{7}(1) + \frac{3}{7}(.91827)$$

$$= 0.96497$$

$$Gain(A3) = E - I(A3)$$

$$= 0.9852 - 0.96497$$

$$\boxed{Gain(A3) = 0.0202} \longrightarrow (3)$$

In conboller, we computated all information gain and observed that all 3 attributes have same value and any one can be randomly selected ~~for~~ as best attribute for splitting. (All this for one first iteration). Also seen how reducer o/p used in computations.