Project II: Alzheimer's Disease diagnosis using patient's gene expression profile

- Feature engineering using MapReduce
- Clustering (Spark)
- Classification (Spark)

- Dataset:

ROSMAP_RNASeq_entrez.csv

ROSMAP_RNASeq_disease_label.csv

patients.csv (optional, additional features of patients)

Part I: Feature Engineering

Clusters

C1: id1, id10, id512
C2: id2, id38
….

ROSMAP

patient	diagnosis	id1	id2	…
p1	1	g(p1,id1)	g(p1,id2)	
p2	3	g(p2,id1)	g(p2,id2)	…
…	…	…	…	…

MapReduce

patient	diagnosis	C1	C2	…
p1	1	g(p1,C1)	g(p1,C2)	
p2	3	g(p2,C1)	g(p2,C2)	…
…	…	…	…	…

$$g(pi, Cj) = \sum_{id_k \in Cj} g(pi, id_k)$$

Part I: Feature Engineering

patient	diagnosis	C1	C2	…
p1	1	g(p1,C1)	g(p1,C2)	
p2	3	g(p2,C1)	g(p2,C2)	…
…	…	…	…	…

MapReduce

Student's T-test

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{S_1}^2}{N_1} + \frac{{S_2}^2}{N_2}}}$$

1. AD
2. NCI

C1	C2	…	Cm
t1	t2		tm

MapReduce

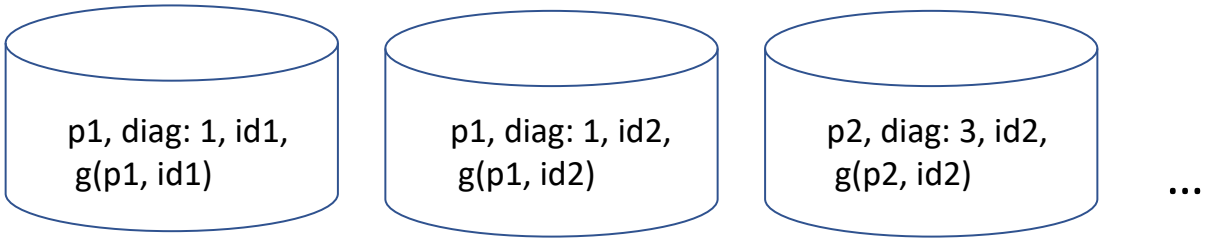
Top-K

C1	C2	…	Cm
t1	t2		tm

The Data Are In HDFS!

patient	diagnosis	id1	id2	…
p1	1	g(p1,id1)	g(p1,id2)	
p2	3	g(p2,id1)	g(p2,id2)	…
…	…	…	…	…

Single processor



HDFS

Algorithm design and pseudo codes

Part II: Machine Learning

- Use **Spark MLlib**
- (**Clustering**) Cluster genes into clusters based on gene expression profile
- (**Classification**)Predict the diagnosis of patients based on their gene expression profile.

Project 2: Part II

- Select top-K clusters as features
- Select a classification algorithm (e.g. Decision Tree, Random Forest) from *Spark Mlib* to train a classification model to predict if a patient has Alzheimer's disease.
- Carry out 3-fold cross-validation

Project II

- Rubric
  - MapReduce Algorithm Design (40 points)
    - Correctness
    - Efficiency
  - Implementation (40 points local + 20 extra) on Amazon EC2)
    - Correctness and efficiency of MapReduce algorithm
    - Usability of your program
  - Written report (20 points)
    - Must present during your project review!
    - Including the following sections
      - Author list
      - Introduction
        - The problem you will solve
      - Methods
        - Detailed description of your algorithm design
        - Detailed description on how to run your program (both local and Amazon)
      - Results
        - Top 10 ranked gene clusters with means and std for AD and NCI, and t-test scores
        - Estimated time complexity vs. number of nodes
      - Discussion
        - The pros and cons of your design
      - Conclusion