EE 542: Internet and Cloud Computing - Project Proposal

Project Proposal	Date: 10/25/18 Academic/Industry Mentor Name: Young Cho E-mail: youngcho@isi.edu	Evaluation criteria: 1.Innovation (15) 2.Final Design Report (25) 3.Completeness(10) 4.Practicality(25) 5.Complexity(10) 6.Design working in H/W(10) 7.All-undergrad team(5)
Team Name/mission statement Trojans / Fight On		

Team Member	e-mail	Role/Responsibility
Karen Wang	wangkh@usc.edu	Team member
Yousef Alaskar	yalaskar@usc.edu	Team member
Abdullah Alhumaymidi	alhumaym@usc.edu	Team member

Problem statement: What problem it solves? Or, what feature does it improve?
Identify top contributing and/or interacting factors for multiple Cancer types, discover genomic and metabolic pathways of tumor growth and metastasis, and develop equivalent but simple classification model leading to better understanding of domain knowledge.

Measure of success: How will you know you have achieved your goal?
Ability to use a small subset of the original features and an interpretable model to classify normal and cancerous tissue subtypes.

Proposal (100-200 words): How will your solution solve the problem described, or improve the feature selected?

Background
Cancer is studied at many levels -- in fact many studies focus on analysis of either upstream DNA genetic variations (i.e. SNP, copy number), or downstream proteomic microRNA gene expression levels. However, it is believed that cancer is caused by a combination of genetic predisposition and epigenetic (gene-environment) factors. For example, although genes are hereditary, DNA methylation can cause genes that were dormant before to be expressed. Silencing of tumor suppressor genes by DNA methylation is a probable mechanism that triggers cancer tumor proliferation.

Hypothesis
Based on the knowledge that cancer can be caused by a combination of factors in multiple levels of the genomic pathway, we will use an integrated approach - examining genome-wide data, including DNA copy number variation and microRNA to build a model that captures the interaction of the data.

Cloud Computing Tools
We will be using AWS SageMaker backed by Apache Spark on AWS Elastic MapReduce (EMR). The Jupyter Notebook instance is connected an Apache Spark cluster running on Amazon EMR. The notebook will interact with Spark clusters via Livy API.

Spark allows distributed data processing, and includes several helpful libraries for machine learning such as MLlib. MLlib supports clustering algorithms (e.g. K-means and Gaussian mixtures), classification algorithms (e.g. logistic regression), Decision trees, random forests, etc. This will accelerate the process of understanding and cleaning-up the data.

SageMaker also contains machine learning algorithms optimized for the hardware they are running on. It claims to deliver up to x10 the performance compared to running the same algorithms elsewhere. This will accelerate the training time of the models.

EE 542: Internet and Cloud Computing - Project Proposal

Last but not least, Jupyter Notebook is the interface to AWS SageMaker. It is user-friendly and allows for fast experiments.

Dataset: GDC database (GDC has 10900 cases that have both miRNA and DNA copy number variation from primary tumor sample type).

Our project can be organized into seven phases:

Phase 1: Build big data pipeline on AWS SageMaker + EMR
- To enable fast processing, we will be using Apache Spark map-reduce on the AWS EMR platform, managed by Hadoop YARN.
- Integrating AWS EMR on SageMaker Notebook https://aws.amazon.com/blogs/machine-learning/build-amazon-sagemaker-notebooks-backed-by-spark-in-amazon-emr/

Phase 2: Understand the data via K-means clustering
- Hierarchical clustering (i.e. divisive, agglomerative)
- Distance measure: pearson correlation, normalized mutual information score

Phase 3: Feature selection
- Use non-linear feature selection algorithms (i.e. XGBoost + recursive feature elimination, kernelized Lasso feature selection etc.) to zero-in on important features.

Phase 4: Build new features, or find optimal way to combine features
- Pairwise pearson correlation coefficients between features from Phase 3
- Topological overlap matrix transformation -- measures node connectivity

Phase 5: Build model
- We aim to use interpretable models (such as Logistic Regression, SVM, Random Forest) to model non-linear dependencies of biomarkers and genes .
- Convolutional Neural Network will most likely give the highest accuracy, but does not provide much insight on the domain knowledge.

Phase 6: Validation
- To overcome sub-optimal amount of data, we do multiple iterations of K-fold cross-validation to ensure that the model isn't overfitted to the training data.
- If time permits, we can utilize the same preprocessing, feature selection and classifier on data of same cancer type from a different database to validate our findings.

Phase 7: Build correlation network
- Present results showing the inferred correlation network of important features

Comparison with Published Results:

Reference	Data	Algorithm	Results
[3]	TCGA DNA methylation profiles for all cancer types	Random Forest, SVM	97%
[2]	Wisconsin Breast Cancer dataset	SVM and RVM hybrid model	96.41%

EE 542: Internet and Cloud Computing - Project Proposal

[4]	TCGA DNA methylation profiles for all cancer types	Convolutional Neural Network	92%

References:

[1] R. Radha, "Using K-Means Clustering Technique To Study Of Breast Cancer," pp. 211–214, 2014.
[2] S. Kumari, "BREAST CANCER CLASSIFICATION USING BIG DATA APPROACH," PARIPEX - INDIAN J. Res., no. January, 2018.
[3] F. Celli, F. Cumbo, and E. Weitschek, "Classification of large DNA methylation datasets for identifying cancer drivers." 2018.
[4] S. Chatterjee, "Convolutional Neural Networks In Classifying Cancer Through DNA Methylation," 2018.
[5] H. Behravan, J. M. Harti, M. Teng, and P. Katri, "Machine learning identifies interacting genetic variants contributing to breast cancer risk : A case study in Finnish cases and controls," no. February, pp. 1–13, 2018.
[6] G. V Glinsky, T. Higashiyama, and A. B. Glinskii, "Classification of Human Breast Cancer Using Gene Expression Profiling as a Component of the Survival Predictor Algorithm," vol. 10, no. 858, pp. 2272–2283, 2004.

	Task 1	**Task 2**
Nov 5th	Build big data pipeline on AWS EMR	Understand the data via K-means clustering
Nov 12th	Feature selection	Build new features, or find optimal way to combine features
Nov 19th	Build model	Validation
Nov 26th	Build correlation network	