Gene of interest recommendation system and logistic prediction model for colorectal carcinoma

Anusha Bellapu
Namratha Shivani Chalasani

INTRODUCTION:

Recommendation Systems (RS) have become important in our data driven digital lives. They play a big role in how we experience information and make choices as consumers. [1] Recommendation Systems (RS) are key in healthcare and data science, tailoring personalized suggestions for effective treatments and enhancing user experiences through precise content recommendations. With a lot of health data around, people are increasingly interested in using RS for selecting genes, especially when dealing with disease data like expression matrices. [2]

A logistic regression model comes into play when we're dealing with a binary outcome. It has been documented that it's easier to model the study outcomes for the better diagnosis if we can consider the logistic regression model and integrate it with RS. [3]

On a worldwide scale, colorectal cancer holds the position of the third most prevalent cancer, making up around 10% of all cancer diagnoses and being the second leading cause of cancer-related deaths globally. It is primarily found in the elderly, with a significant number of cases occurring in individuals aged 50 and older. Advanced stages often characterize the identification of colorectal cancer, resulting in restricted treatment alternatives. It's crucial to discover gene markers for colon cancer to decrease fatalities and improve early detection. Implementing RS would speed up and refine the identification of relevant genes, greatly enhancing targeted therapy for colon cancer. [4]

RESEARCH PROBLEM:

This study focuses on the significant challenge of colorectal cancer, a global health issue. Despite ranking as the third most common cancer globally, colorectal cancer persists as a major cause of mortality. Many cases go unnoticed until advanced stages, restricting treatment choices. Hence, there is a critical need to pinpoint precise gene markers linked to colon cancer for early detection and improved patient outcomes.

The emergence of precision medicine, customizing treatments for individual patients, has shown effectiveness in managing and curing conditions. In this study, our aim is to develop an algorithm for identifying top genes of interest and a prediction model. This approach enables personalized therapy targeting specific genes for each patient.

METHODS:

Using information obtained from the TCGA-COAD project [5], consisting of RNA-seq gene expression data from 481 primary tumor samples and 41 normal samples. Transcriptome profiling of normal and tumor samples in TCGA-COAD was acquired using the TCGAbiolinks R package [6]. Samples with missing metadata were excluded, and the DESeq2 R package [7] was employed for performing differential gene expression analysis to identify highly differential genes for further investigation.

The dataset is divided into training and testing sets using an 80:20 ratio. The system calculates and recommends the top N genes of interest to improve cancer detection precision and facilitate more focused therapeutic interventions. The RS algorithm evaluates gene precision and coverage to assess the effectiveness of gene recommendations, while the logistic regression model for diagnosis shows potential for progress in the early detection and treatment of colorectal cancer.

EXPERIMENTAL DESIGN:

Gene interest calculation:

The gene expression data is log-transformed, and the top 1000 genes are selected and stored in a matrix form (dual mode matrix). A single gene network (G) is constructed by cross-multiplication

of the gene expression matrix (Ge) with itself. A similarity matrix (Ja) [8] is constructed and used to generate the gene of interest (Gi) by matrix multiplication of the similarity matrix with the gene expression matrix. The gene of interest is then sorted to get the gene ranking (GR). [9]

Gene single network = gene expression matrix * gene expression matrix

Gene similarity matrix (Ja): $Ja_{ij} = \frac{gij}{gii+gjj-gij}$

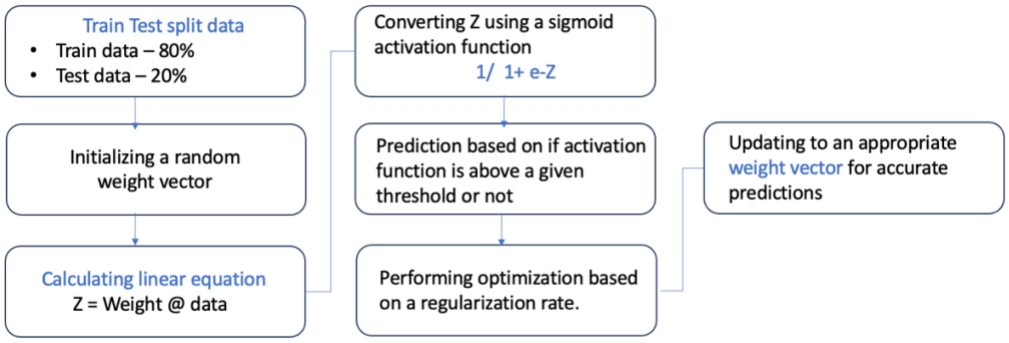
Gene interest (Gi) = Gene similarity matrix * gene expression matrix

Gene Rank (GR) = sort Gene of interest

Logistic Regression Model:

A logistic regression model has been employed from scratch to predict the patient diagnosis from the gene expression data and covariates like age, stage of tumor and sex. Data preprocessing was performed to convert the categorical variables like stage of tumor, sex, and diagnosis into numerical indices.

Model:



The Python sklearn package [10] was utilized to split the data into training and test sets, with 80% allocated to the training data and 20% to the test data. Initially, a random weight vector was generated using the np.random function [11], and this vector was employed to make the initial predictions prior to optimization.

The data was linearly fitted by multiplying the weight vector with the x data and then transformed using the sigmoid activation function. Predictions were made based on whether the transformed data exceeded a threshold of 0.5 or not.

The weight vector was optimized to minimize the loss of predictions, utilizing a regularization rate of 0.0001 and learning rate of 0.01 over 100 epochs of iterations. Subsequently, data predictions were conducted again to evaluate how the optimization affected the accuracy of predictions.

EXPERIMENTAL RESULTS:

Recommendation system: The gene single network creation resulted in a gene * gene matrix. This matrix is converted into similarity with diagonals as 1 and the higher the score more similar the genes are. Gene of interest matrix is a gene * sample matrix which has the values of each gene for the sample based on their similarity. These are then ranked to get the top genes.

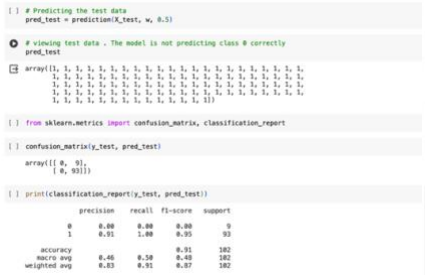
```
# getting the top genes for one sample in the coad data
rec = recommendation(pd.DataFrame(coad_data[coad_data.columns[2]]))
```

```
rec
```

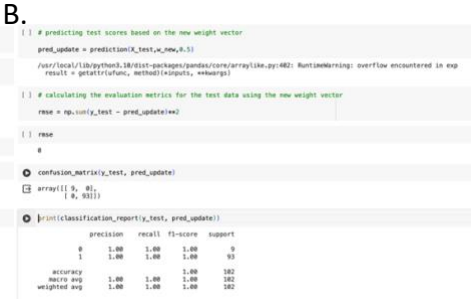
```
{'TCGA-D5-5538-01A-01R-1653-07': Index(['ENSG00000166126.11', 'ENSG00000156381.9', 'ENSG00000149809.16',
        'ENSG00000107833.10', 'ENSG00000138031.14', 'ENSG00000161888.11',
        'ENSG00000012171.20', 'ENSG00000135540.11', 'ENSG00000143248.13',
        'ENSG00000125144.14'],
       dtype='object')}
```

Logistic regression: The model created with sample weights that are generated at random showed that the matrix is not predicting the class 0 with all class 0 misclassified as class 1. To correct this optimization with 0.0001 regularization and 0.01 learning rate method showed that the model was able to predict all the data with 100% accuracy.

A.



B.



A. Predictions before weight vector optimization.
B. Predictions after weight vector optimization.

DISCUSSION ABOUT RESULTS:

This study focuses on integrating Recommendation Systems (RS) and logistic regression to address the critical issue of colorectal cancer. RS, rooted in big data, offers personalized gene recommendations, an asset in the era of precision medicine. The logistic regression model complements this approach by predicting patient diagnoses, facilitating targeted therapeutic interventions.

The pressing concern of colorectal cancer's impact is highlighted, emphasizing its global prevalence and the need for early detection strategies. The RS algorithm, using RNA-seq data from the TCGA-COAD project, computes top gene recommendations. These genes are vital in understanding and addressing colorectal cancer at a molecular level.

The gene interest calculation method involves log transformation, gene network construction, and similarity matrix generation. This process yields a ranked list of genes, providing a foundation for the subsequent logistic regression model. The logistic regression model, initially hindered by misclassifications, is successfully optimized with regularization, and learning rate adjustments, achieving a remarkable 100% accuracy in predicting patient diagnoses.

CONCLUSION:

The integration of RS and logistic regression presents a promising avenue for advancing precision medicine in colorectal cancer. By combining personalized gene recommendations with accurate prediction models, this study lays the groundwork for tailored therapeutic strategies. The ability to compute top genes of interest ensures a targeted approach to cancer treatment, a crucial step in managing and potentially curing colorectal cancer.

Moving forward, the study's findings offer insights into the potential clinical applications of RS and logistic regression in the realm of cancer research. The success in optimizing the logistic regression model signifies a robust framework for enhancing diagnostic accuracy. These methodologies, when applied synergistically, contribute to a more effective and personalized approach to colorectal cancer management, holding implications for broader applications in precision medicine.

References:

1. Fayyaz Z, Ebrahimian M, Nawara D, Ibrahim A, & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences*, *10*(21), 7748.
2. Chang V. Towards data analysis for weather cloud computing. (2017). *Knowl Based Syst*.
3. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic regression in clinical studies. International Journal of Radiation Oncology* Biology* Physics. 2022 Feb 1;112(2):271-7.
4. World Health Organization. (2023, July 11). Colorectal cancer. **https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer.**
5. National Cancer Institute. (2019). The Cancer Genome Atlas - Colon Adenocarcinoma (TCGA-COAD) Data. *Genomic Data Commons*. https://portal.gdc.cancer.gov/projects/TCGA-COAD .
6. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research. 2016 May 5;44(8):e71-.
7. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014 Dec;15(12):1-21.
8. Vorontsov IE, Kuakovskiy IV, Makeev Vj. Jaccard index-based similarity measure to compare transcription factor binding site models. *Algo Mole Biol,* 2013;8(1)1.
9. Hu J, Sharma S, Gao Z, & Chang V. (2018). Gene-based collaborative filtering using recommender system. *Computers & Electrical Engineering*, *65*, 332-341.
10. Pedregosa F, Varoquaux, Ga el, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*. 2011; 12:2825–30.
11. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020; 585:357–62.