11/17 Final Project Interim Update

Data Procurement and Preprocessing: I have applied for and downloaded the November 2019 version of the SEER dataset for thyroid cancer.

Review of the Literature: Keeping with my proposed timeline for this project, I completed a literature review on related work in prediction of cancer outcomes using machine learning techniques. Commonly used methods include decision trees, logistic regression, artificial neural networks, and support vector machines.[1, 2] Ensemble methods have also been tried with good success in colorectal cancer SEER data, with a significant performance improvement attributable to the synthetic minority over-sampling technique (SMOTE).[3] Recent work has examined prostate, cervical, breast, lung and thyroid cancer.[4, 5] On larger datasets, ANN appear to have superior performance to other machine learning methods; on smaller datasets, certain algorithms such as SVM sometimes outperformed ANN. [6, 7]

Implementation Plan

Preprocessing: for thyroid cancer, class imbalance has consistently presented a challenge, as five year survival for the most common subtypes are extremely high. To address this, I plan to implement the SMOTE technique to generate synthetic samples of the minority class. For missing data, *k*-nearest neighbors will be used to impute missing values. Discrete features will be handled using a one-hot encoding.

Feature selection: a number of key features in the SEER dataset will be selected to start, and this number may be reduced using regularized regression, or principal components analysis (PCA) as part of a pre-processing step before ANN to prevent overfitting. These features will include age at diagnosis, grade, American Joint Committee on Cancer (AJCC) T, N and M stage values, primary tumor site, positive lymph nodes, and tumor size.

Measures for performance evaluation: for each classifier, area under the receiver operating characteristic (AUROC) will be reported along with a confidence bound using bootstrapped data samples. This will allow for direct comparison between classifiers.

Algorithm update: rather than using PCA for classification (an application to which it is not well suited), the second algorithm examined will be the support vector machine (SVM) with the preprocessing steps described above.

Next Steps

I will follow the original timeline and proceed with coding, debugging, and testing implementations of the algorithms by 12/1/2020.

References

1. Wen, H.; Li, S.; Li, W.; Li, J.; Yin, C. In *Comparision of Four Machine Learning Techniques for the Prediction of Prostate Cancer Survivability*, 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 14-16 Dec. 2018; 2018; pp 112-116.
2. Mourad, M.; Moubayed, S.; Dezube, A.; Mourad, Y.; Park, K.; Torreblanca-Zanca, A.; Torrecilla, J. S.; Cancilla, J. C.; Wang, J., Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. *Scientific Reports* **2020,** *10* (1), 5176.
3. Al-Bahrani, R.; Agrawal, A.; Choudhary, A. In *Colon cancer survival prediction using ensemble data mining on SEER data*, 2013 IEEE International Conference on Big Data, 6-9 Oct. 2013; 2013; pp 9-16.
4. Tseng, C.-J.; Lu, C.-J.; Chang, C.-C.; Chen, G.-D., Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Computing and Applications* **2014**.
5. Das, A.; Ngamruengphong, S., 81 Machine Learning Based Predictive Models Are More Accurate Than TNM Staging in Predicting Survival in Patients With Pancreatic Cancer. *Official journal of the American College of Gastroenterology | ACG* **2019,** *114*, S48.
6. Kim, W.; Kim, K. S.; Lee, J. E.; Noh, D.-Y.; Kim, S.-W.; Jung, Y. S.; Park, M. Y.; Park, R. W., Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of breast cancer* **2012,** *15* (2), 230-238.
7. Cirkovic, B. R. A.; Cvetkovic, A. M.; Ninkovic, S. M.; Filipovic, N. D. In *Prediction models for estimation of survival rate and relapse for breast cancer patients*, 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), 2-4 Nov. 2015; 2015; pp 1-6.