Nicolas Vuille-dit-Bille

Avenue de Morges 20

1004 Lausanne

nicolas.vuille-dit-bille@agroscope.admin.ch; vuillenicolas@gmail.com

The impact of molecular profiling of gliomas on treatment and prognosis - Galina Glousker

Peer consulting report

30 December 2024

Project description

For her data science project, Galina Glousker investigated the impact of molecular profiling of gliomas on treatment and prognosis (https://github.com/GalinaGl/CAS_ADS_2024/tree/main/Module3). She included in her analysis two main category of adult -type diffuse glioma according to the isocitrate dehydrogenase (IDH) mutation status: wild type and mutated. Presence or absence of different types of gliomas can lead to different prognosis of tumors (e.g. Glioblastoma, Astrocytoma and oligodendroglioma). The project consisted of three main parts: analyzing genomic data from adult glioma patients using available databases, identifying genomic alterations linked to clinical behavior and therapy outcomes, and applying machine learning techniques to predict glioma outcomes based on tumor molecular profiles. The collected data were described according to different features like sex distribution, neoplasm histologic type and grade distribution.

Data processing

Two main data sources were used for this project: sequencing and clinical data. The sequencing information was firstly filtered to identify silent mutations and top 20 mutated genes. Then the sequencing data table was merged with clinical information including the following features: 'Patient ID', 'Diagnosis Age', 'Sex', 'Neoplasm Histologic Grade', 'TMB (nonsynonymous)' and 'Pan-Glioma DNA Methylation Cluster'. At the end, some NaN were created after the merge but they were dropped before further data analysis. At this stage, it was not completely clear why some NaN were created in the merging process. One possible improvement could be to look more in details at the results of the merged dataset to avoid excluding relevant information for the study case when dropping the NaN.

The final merged data was then explored by using histogram visualization. There seemed to be no notable skewness, such as unbalanced sex and diagnosis age distribution, that could impact further analysis. This was a good approach of data exploration to understand better the problems that could occurred with machine learning models.

Unsupervised learning

Unsupervised learning approach was used to explore more deeply the data structure and to perform dimensionality reduction. The PCA showed that between 11 and 12 principal components explained 90% of the variance. Then hierarchical clustering was performed on top of PCA suggesting seven optimal clusters to characterize the dataset. Gaussian mixture model (GMM) clustering on UMAP-reduced data also confirmed seven optimal clustering group (GMM Silhouette Score: 0.61 and Davies-Bouldin Index: 0.56). According to the unsupervised learning results and visualization, there was a good investigation showing that at the end one specific feature, the diagnosis age, seemed to have a high impact on variance distribution and on the clusters organization. In conclusion, various unsupervised learning methods were applied with success to better understand the data and reduce its dimensionality. However, there seemed to have some confusion about the number of clusters when reproducing the results in the jupyter notebook. Some codes might have been edited without being specifically annotated. Adding more detailed annotations to the edited sections would enhance reproducibility and save time when revisiting the code after a period of inactivity.

Supervised learning

Supervised learning methods were investigated to perform molecular diagnostics and survival predictions. This analysis part referred to a publication using a hierarchical voting stages-based ensemble learning scheme (Tasci et al. 2022). This approach combined methods such as Weight of evidence, Recursive feature elimination, Random Forest, and LASSO to derive a reduced feature set. The resulting features were then utilized in predictive models, including Logistic regression, Support vector machines, K-nearest neighbors, Random Forest, and Adaboost, to classify gliomas into two categories: low-grade gliomas (LGG) and high-grade gliomas with glioblastoma multiforme (GBM). These methods were successfully implemented in the case of study. It was interesting to see the different results derived from different supervised learning methods. At the end there was still common pattern comparing the methods. For example, both Adaboost and support vector machine highlighted the importance of diagnosis age in the predictions but not with the same substantial impact. To understand the outputs of the models and compare them with the goal of the study (molecular diagnostics and survival prediction), the results were visualized efficiently in classification reports including precision, recall, f1-score and support. This allowed to compare easily the different outputs with relevant conclusion.

References and Bibliography

Galina Glousker. 2024. CAS_ADS_2024. https://github.com/GalinaGl/CAS_ADS_2024/tree/main/Module3. (2024).

Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., & Krauze, A. V. (2022). Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. International Journal of Molecular Sciences, 23(22), 14155.