AOS C204 Project

Unsupervised Classification of Traumatic Brain Injury Severity using Kmodes Machine Learning Method

Parnian Hemmati

UID: 205945856

December 8, 2023

Abstract

Traumatic Brain Injury (TBI) is a concerning public health problem. TBI groups into three main clusters as mild TBI, moderate TBI and severe TBI. This classification of TBI is of importance as it indicates treatment and post injury care. In patients with specific occupations such as professional athletes it is even more important to indicate TBI severity and based on that predict the length of athlete rest from career. However, it is really challenging to cluster TBI, as it depends on lots of clinical, neuroimaging and genetic factors. In this project, an unsupervised machine learning method is used to cluster patients in three different groups based on TBI severity. To make this possible, a rich dataset provided by the NIH/NINDS Traumatic Brain Injury Common Data Elements (TBI-CDEs) initiative is utilized. This dataset includes data from 586 patients focusing on genetics, injury metrics, neuroimaging factors and long-term markers. 12 features of this dataset are used to train the unsupervised kmodes model. Elbow plot for different featuresets and cost ratios are extracted to showcase the influence of each feature on clustering. To gain more insight into the clusters, 4 long-term factors are studied for each cluster and compared. These factors include Glasgow Outcome Scale-Extended (GOSE) 3 and 6 months post injury. Statistical analysis shows one of the groups noticeable has higher GOSE score both 3 and 6 months post injury; suggesting the clusters can be indicators of TBI severity. Further, two other indicators as Wechsler Adult Intelligence Scale (WAIS) and Post-Traumatic Stress Disorder (PTSD) are studied. Regarding WAIS score, the average values in group 2 is meaningfully higher, supporting the TBI severity based clusters. However, for PTSD, all the clusters show same trend as more patients with no PTSD. This arises the question of whether PTSD can be a good indicator of TBI severity as it is not influenced only by physical damage, but also emotional trauma as well. Given all these, this project shows that an unsupervised clustering of TBI severity is possible, however it needs further investigation and more extended use of available methods. A preliminary investigation of features influence on clustering, suggests that that skull fracture, and Rotterdom computed tomography (CT) score can be used as TBI severity indicators while CT intracranial pressure does not contribute significantly to this clustering. . It also shows that GOSE score 3 and 6 months post injury as well as WAIS score can be used as TBI severity indicators.

1 Introduction

Traumatic brain injury (TBI) is a critical public health concern, as it can lead to diverse outcomes that require tailored treatment and care. TBI is of importance considering its profound impact on brain function, both in short and long term. In the united states, TBI contributes to approximately 30% of injury-related death [1]. TBI is a complex disorder that is traditionally stratified based on clinical signs and symptoms. Recent imaging and molecular biomarker innovations provide unprecedented opportunities for improved TBI precision medicine, incorporating patho-anatomical and molecular mechanisms [2].The complexity of TBI severity classification arises from the multifaceted nature of the condition, involving a wide range of clinical, imaging, and genetic parameters, making precise prediction a challenging yet vital task for improved patient management.

This project utilizes an unsupervised machine learning approach, specifically the K-Modes algorithm, to cluster TBI cases based on severity. By using a rich dataset including clinical, imaging, and genetic parameters, this research aims to study patterns, relationships, and potential biomarkers that can enhance our understanding of TBI severity and pave the way for more targeted and effective interventions. A comprehensive dataset from the NIH/NINDS Traumatic Brain Injury Common Data Elements (TBI-CDEs) initiative is used for this purpose. Utilizing unsupervised classification, this project seeks to contribute valuable insights to traumatic brain injury research and clinical management as for indicating biomarkers that contribute the most in severity classification and developinf methods to predict long-term side effects of TBI.

2 Data

The dataset used in this study originates from the NIH/NINDS-developed Traumatic Brain Injury Common Data Elements (TBI-CDEs) initiative. To address challenges in TBI clinical research, such as data standardization, patient stratification, and variability in injury types, the TBI-CDEs focused on four major domains: clinical assessments and demographic information, genetics and proteomics, neuroimaging, and outcome measures. The dataset follows

an approach as categorizing data elements into 'core' (fundamental information like gender and age), 'basic' (additional diagnostic details such as education level and cause of injury), and 'supplemental' (including emerging elements like imaging and serial plasma biomarkers) categories. The multicenter prospective TRACK-TBI Pilot study, involving a limited 3-center clinical observational trial, assessed the feasibility and utility of TBI-CDEs. The dataset includes essential features such as CT Brain Pathology, alongside genetic information (COMT), cognitive scores, and post-injury GOSE scores. This dataset is made available by a NIH Traumatic Brain Injury Research Center program.

Variable	N	Missing	Min	Max	Mean	SD
CT Brain Pathology	586	0	0	1	0.44	0.50
Skull Fracture	586	0	0	1	0.22	0.41
Skull Base Fracture	586	0	0	1	0.11	0.31
Facial Fracture	586	0	0	1	0.17	0.38
Epidural Hematoma	586	0	0	1	0.05	0.22
Subdural Hematoma	586	0	0	1	0.26	0.44
Subarachnoid Hemorrhage	586	0	0	1	0.26	0.44
Contusion	586	0	0	1	0.24	0.43
Midline Shift	586	0	0	1	0.07	0.25
Cisternal Compression	586	0	0	1	0.12	0.33
Marshall CT Score	586	0	1	6	1.76	1.10
Rotterdam CT Score	586	0	1	6	2.45	0.83
PTSD Diagnosis at 6 months (DSM-IV)	338	248	0	1	0.24	0.43
PTSD Checklist-Civilian Version at 6 months	338	248	17	83	32.98	14.80
WAIS Processing Speed at 6 months	305	281	50	150	99.20	15.96
CVLT: Short Delay Cued Recall at 6 months	296	290	-4.0	2.5	-0.08	1.14
CVLT: Long Delay Cued Recall at 6 months	295	291	-3.5	2.5	-0.19	1.17

3 Modeling

In order to cluster patients based on traumatic brain injury severity in the unlabeled dataset, an unsupervised clustering approach was taken. Given that the dataset includes distinct and categorical features related to TBI for hospitalized patients, the choice of the "Kmodes" clustering algorithm seemed appropriate. This selection aligns with the categorical nature of the features, as "Kmodes" is an appropriate method to study non-continuous and distinctly

numbered datasets.

To initiate the clustering process, the objective was to categorize TBI severity into three distinct groups: mild, moderate, and severe. With this goal in mind, the number of clusters (n) was set to 3 to align with the severity levels. The clustering model was trained using the "Kmodes" algorithm, and subsequent analyses were performed to evaluate the effectiveness of the clustering solution in accurately stratifying patients based on TBI severity. The choice of "Kmodes" matches the nature of this dataset and allows the exploration of categorical feature patterns.

4 Results and Discussion

4.1 Clustering and Feature Selection

For the clustering analysis of the dataset, twelve features were initially selected. These features were all injury metrics, factors that conventionally are used to indicate TBI severity. The elbow plot, shown in Fig.1, is plotted and used to determine the optimal number of clusters. The plot suggests that a choice between k=3 or k=4 is suitable for clustering, providing a clear inflection point and cost gradient. To investigate the relationships between

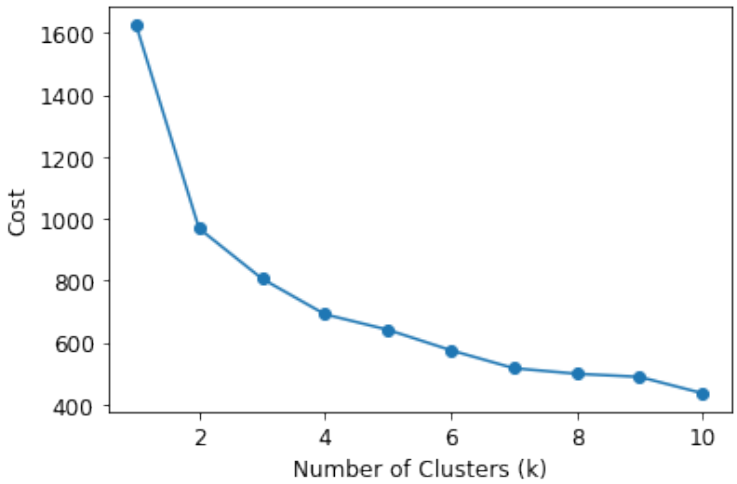

Figure 1: Elbow plot for the complete dataset.

different factors and injury metrics contributing to clustering, two distinctive methods were employed. Firstly, an elbow plot was generated for the dataset excluding one feature at a

time. Fig.2a illustrates how these elbow plots evolve with the exclusion of each factor. This methodology offers valuable insights into the interplay and correlations among features, showcasing that no single feature dominantly influences clustering. The absence of a feature does not yield a substantial change in the elbow plot, highlighting the interdependence of the considered features. Furthermore, to quantify the contribution of each feature set to

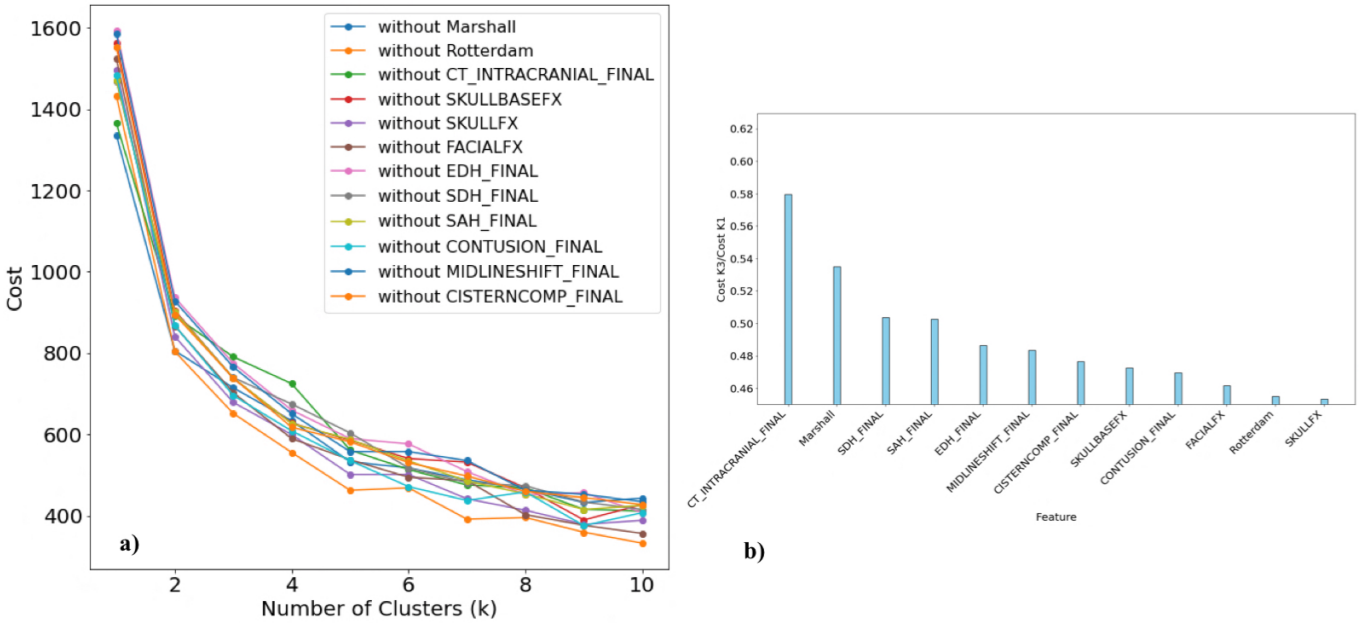

Figure 2: Features contribution to clustering a)Elbow plot for different feature sets. b)Cost ratio for one cluster and three clusters for each feature set.

clustering, another approach was implemented. The ratio of the cost for clustering in three groups to the cost of clustering in one group was calculated for each feature set. A lower ratio indicates a reduced cost for three groups, signifying a more optimal feature set for clustering. Fig.**??** visually represents this calculated ratio, providing an overview of the effectiveness of each feature set in the clustering process. It can be seen that skull fracture and Rotterdom computed tomography (CT) score are two bold biomarkers of TBI severity, followed by facial fracture and cisternal compression. On the other hand, CT intracranial pressure contribution to clustering is smaller compared to other features.

These two methodologies, as feature exclusion analysis and ratio-based assessment, prive a better understanding of different features and markers contribution in TBI severity identification. It also provides more insight to understanding of the dataset's inherent structures

and the dynamic roles played by individual features in the clustering.

4.2 Analysis of Post-Impact Long-Term Indicators in Clustered TBI Severity

In order to investigate the representativeness of the identified clusters for TBI severity, an insightful exploration into post-impact long-term indicators was conducted. Noteworthy indicators such as PTSD Diagnosis at 6 months (DSM-IV), WAIS Processing Speed at 6 months,Glasgow Outcome Scale-Extended (GOSE) after 3 months, and GOSE after 6 months were selected for this study.

It is important to highlight that these particular features were not utilized during the initial clustering phase. Post-model training, statistical analyses were performed to study whether notable distinctions among the clusters could be observed. We aim for studying whether the identified clusters can be potential representative markers of TBI severity.

First parameters to study were GOSE 3 and 6 months post injury. Fig **??** illustrates GOSE scores after 3 and 6 months for patients in each group. It can bee seen that group 2 noticeably includes higher GOSE scores, which supports the hypothesis that the clusters can be representatives of TBI severity specifically with group 2 showcasing severe cases of TBI compared to group 3 as mild TBI. It can be seen that higher GOSE scores are reported noticeably more in patients of group 2. WAIS processing speed at 6 months. Fig. 4 shows average WAIS score for each cluster. It can bee seen that group 2 represents noticeably higher value of average WAIS score, followed by group 1 and group 3 with the lowest average. This shows that unsupervised clustering have succeeded in identifying one group of patients with severe TBI. Eventually, PTSD for each cluster was studied. As shown in Fig. 4, the current clustering does not represent a valid way to predict PTSD since patients in all of the three clusters mostly reported no PTSD 6 months post injury. However group 2 includes most of the patients who still reported PTSD after 6 months, it also includes most of the patients who reported negative PTSD. All of the three groups include more patients reporting no PTSD which means that the underlying factors that resulted in the three different clusters, does not behave the same with PTSD. It also brings up the question

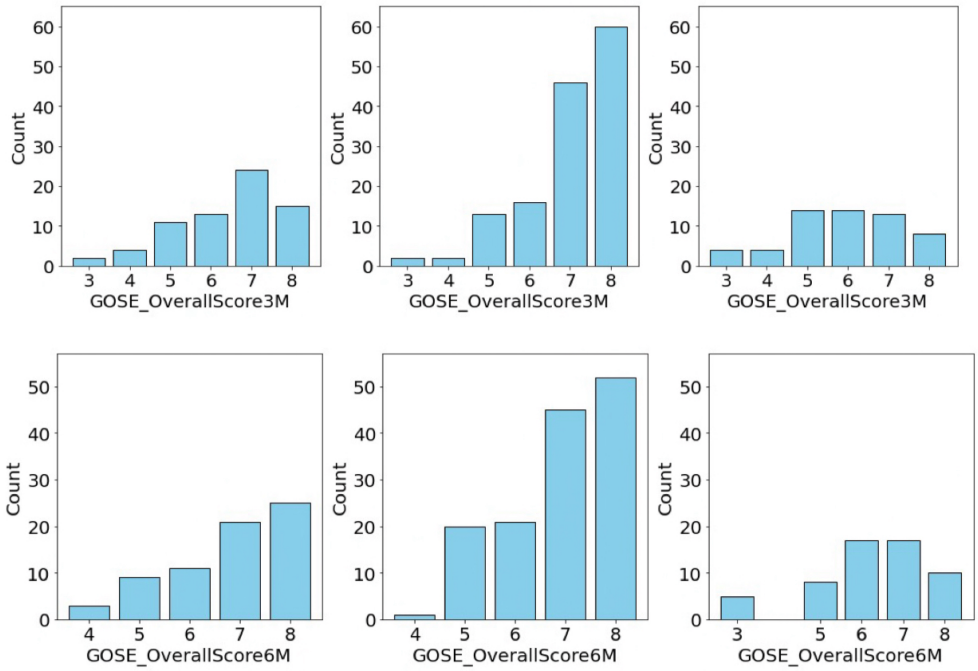

Figure 3: GOSE scores 3 months and 6 months post injury.

that whether PTSD 6 months post injury can be a good TBI severity indicator, as the individual post traumatic stress disorder may not only come from the severity of physical damage but also emotional trauma as well. Studying these long-term indicators within each cluster, we seek to unravel any discernible differences that may serve as indicative elements of TBI severity. This comprehensive approach, incorporating both unsupervised clustering and subsequent analysis of key long-term indicators, contributes to the understanding of the relationship between TBI severity and post-impact outcomes.

5 Conclusion

In conclusion, this project employed the K-Modes unsupervised machine learning algorithm to cluster traumatic brain injury (TBI) cases based on severity using a diverse dataset encompassing clinical, imaging, and genetic parameters. The analysis showed three distinct clusters representing potential clusters of TBI severity. Feature selection analyses highlighted the interdependence of various factors in the clustering process, and subsequent exploration of post-impact long-term indicators within these clusters provided valuable insights into

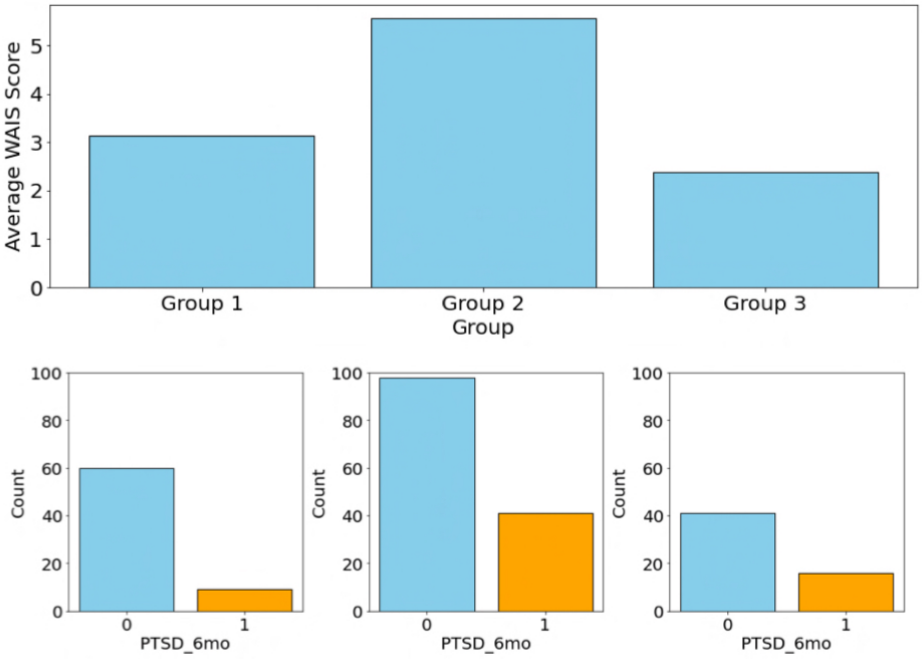

Figure 4: Average WAIS score and PTSD for each group.

their representativeness. Notably, the study identified a cluster with higher severity, as indicated by cognitive scores and Glasgow Outcome Scale-Extended GOSE scores. However, the clustering did not effectively predict post-traumatic stress disorder (PTSD) 6 months post-injury, suggesting that emotional trauma might contribute independently to PTSD. This project showed that TBI severity can be studied using machine learning methods, gaining insight into biomarkers and long-term effects of severe TBI. To expand this project further, a more comprehensive study on factors can be done, including principle component analysis to decrease the dimentionality od the system for clustering. Moreover, clustering method can be developed by studyinbg different feature sets coming from 12 features and indicating the optimum feature set for clustering. Later, by means of long-term identicors, this model can be used to predict the long-term effects of TBI based on its severity, which can be really helpful in clinical perspective.

References

[1] J. Lang, R. Nathan, D. Zhou, X. Zhang, B. Li, and Q. Wu, "Cavitation causes brain injury," *Physics of Fluids*, vol. 33, Mar. 2021.

[2] J. L. Nielson, S. R. Cooper, J. K. Yue, M. D. Sorani, T. Inoue, E. L. Yuh, P. Mukherjee, T. C. Petrossian, J. Paquette, P. Y. Lum, G. E. Carlsson, M. J. Vassar, H. F. Lingsma, W. A. Gordon, A. B. Valadka, D. O. Okonkwo, G. T. Manley, and A. R. Ferguson, "Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis," *PLOS ONE*, vol. 12, p. e0169490, Mar. 2017.