Glioma Classification Project – SVM Discussion

Mitchell Otley **(23475725)**
James Wigfield **(23334375)**
Nate Trew **(23120643)**

Data Collection

The data used to train the SVM is synthesized from the extracted conventional features (maximum tumor area, maximum tumor diameter and outer layer involvement), and our top ten intensity-based, shape-based, and texture-based radiomic features of each MRI volume in the dataset. The SVM model concatenates this data with data extracted from the *name_mapping.csv* file, specifying the classification of each volume as to being High-Grade Glioma (HGG) or Low-Grade Glioma (LGG).

The result is a csv table of the dimension 369x35 for the provided dataset (33 columns for the extracted features of each of the 369 MRI volumes, one column for the volume label and one column for the grading of the tumor).

	A	B	C	D	E	F	G	H	I	J
1	BraTS_2020_subject_ID	Grade	area	diameter	out_layer_involvement	MinorAxisLength3D	LeastAxisLength3D	VolumeDensityAEE_3D	VolumeMesh3D	Elongation3D
2	volume_001	HGG	5048	104.8	4.5	73.7959879	55.72137928	1.152807593	211768	0.864932974
3	volume_002	HGG	2105	85.4	2.2	49.77191045	37.15595469	1.093291521	66779	0.789026808
4	volume_003	HGG	942	55.5	1	40.26938618	28.65070947	0.798992217	29657.75	0.655372735
5	volume_004	HGG	2751	78.1	3.7	53.83290369	45.57621767	1.216546416	103392.5	0.81371467
6	volume_005	HGG	779	63.1	2.2	32.90797638	28.47553939	0.61438638	21863	0.453737602
7	volume_006	HGG	3179	96.9	3	59.46106879	46.00336172	1.119730592	138936	0.686360608
8	volume_007	HGG	1501	67.3	0.9	42.1783324	27.66262668	1.13302052	39588	0.737471189
9	volume_008	HGG	1272	71.3	0.8	39.28865146	29.64257243	0.951499879	33344	0.683659391
10	volume_009	HGG	4219	120.6	6.4	66.223175	46.97766356	1.107228279	182749.75	0.653567431
11	volume_010	HGG	1331	77	2.8	46.11753052	34.23128893	0.773081839	44188	0.666922079
12	volume_011	HGG	1377	66.1	1.1	46.09098207	35.1005393	0.932162821	54950	0.662320852
13	volume_012	HGG	1399	65.6	1.8	37.65019681	32.87958228	0.864927232	31770	0.664388009
14	volume_013	HGG	1632	75	0	43.16602873	41.91116574	0.921860039	46092	0.817809201
15	volume_014	HGG	2372	92.1	1.9	50.35246631	37.80559019	1.085421801	80337	0.678077071
16	volume_015	HGG	2794	86.3	0.1	62.50647996	42.03642882	1.176840782	125138	0.808728986
17	volume_016	HGG	3024	89.4	3	64.40451645	47.84098087	1.084991813	126976	0.887842956
18	volume_017	HGG	2299	69.9	2.3	53.08988878	45.79019452	1.045885563	100570	0.702765507
19	volume_018	HGG	1689	75.2	0	52.73238598	38.63616414	0.481885582	44851	0.604392779
20	volume_019	HGG	1372	59.5	3.7	39.68173275	33.60296043	1.24051106	43522	0.78967649
21	volume_020	HGG	4223	125.1	6.7	65.06182364	46.62864829	1.080225587	175445.5	0.636320816
22	volume_021	HGG	1597	80.3	0.7	45.95356368	34.42050914	0.574730933	36546	0.598520515
23	volume_022	HGG	1889	61.1	0	44.90092414	42.49136238	1.206540227	67995	0.795929642
24	volume_023	HGG	1316	67.7	0.1	43.84548839	34.52589942	0.904682994	40982	0.767179028
25	volume_024	HGG	1935	65.7	2	50.82525583	43.55981278	0.805363774	95312	0.497838276
26	volume_025	HGG	2028	88.8	0.9	51.95037006	43.56311661	1.040576696	73302	0.873882644
27	volume_026	HGG	3878	109	1.2	64.33231163	51.3430381	1.107615948	180475	0.682826919
28	volume_027	HGG	1818	77.7	1.6	49.14144518	32.07654569	0.999268115	66320.75	0.611103328
29	volume_028	HGG	1071	58.9	0	31.25376195	24.29536737	1.203937054	20760	0.720614721
30	volume_029	HGG	1898	88.7	1.3	42.45400226	40.32651863	0.936950564	62205	0.573215601
31	volume_030	HGG	2137	81.7	1.4	61.55337074	50.16634968	0.815518022	99266	0.817613082

1: Sample of combined_features.csv

The conventional features and radiomic features are all obtained through the 'Extract Conventional Features' and 'Extract Radiomic Features' buttons in our MATLAB GUI.

Feature Selection

Conventional Features:

- Maximum Tumor Area
- Maximum Tumor Diameter
- Outer Layer Involvement

Intensity-Based Radiomic Features:

- Discretised Intensity Skewness
- Intensity Kurtosis
- Minimum Histogram Gradient
- Minimum Discretised Intensity
- Volume at 10% Intensity Fraction
- Volume Fraction Difference between Intensity Fractions
- Discretised Intensity Entropy
- Maximum Intensity
- Intensity Histogram Coefficient of Variation
- Global Intensity Peak

Shape-Based Radiomic Features:

- Minor Axis Length
- Smallest Axis Length
- Volume Density (Approximate Enclosing Ellipsoid)
- Volume (Mesh)
- Elongation
- Volume (Voxel Count)
- Major Axis Length
- Flatness
- Surface to Volume Ratio
- Spherical Disproportion

Texture-Based Radiomic Features:

- Information Correlation 1 - Merged
- Information Correlation 1 - Averaged
- Dependence Count Percentage
- Normalised Inverse Difference Moment - Merged
- Normalised Inverse Difference Moment - Averaged
- Normalised Inverse Difference - Merged
- Normalised Inverse Difference - Averaged
- Information Correlation 2 - Merged
- Information Correlation 2 - Averaged
- Run Entropy - Merged

Data Partitioning

As per the project specifications, 10 randomly selected HGG and LGG patients are assigned to a 'hidden' testing set. The remaining 349 patients are utilised in the training and validation of the SVM. However, of these 349 patients, only 66 patients are classified as having LGG. As the SVM is sensitive to unbalanced datasets, the resulting model would exhibit a bias towards classification of unseen samples to the class with a larger representation (in this case, HGG). To mitigate this, we reduce the dataset passed to the SVM to be equal between the binary classes. In essence:

- 10 LGG and 10 HGG patients assigned to testing set
- 66 LGG Patients assigned to training set
- 66 HGG Patients assigned to training set
- 217 HGG Patients unassigned (not used in training or testing of the SVM)

We don't want 217 data samples being unused, so to account for this we train 1000 iterations of the SVM Classifier, each time using a different random sampling of 66 HGG patients of the available 283. We then select the model with the highest training accuracy as our preferred model to use on the testing set.

During the model training phase, we utilise *k-fold cross validation* to create a validation set. This method splits the training data into *k=5* same-sized groups in our model, wherein each group is used to perform a validation of the other groups of data in the training set. We chose *k=5* based on empirical recommendations.[1]

Model Accuracy

We opted to use MATLAB's Classification Learner App to develop a model, and exported it to a function to incorporate with our data.

```
>> svm_train
New Highest Accuracy: 0.65152(Iteration 1)
New Highest Accuracy: 0.68182(Iteration 3)
New Highest Accuracy: 0.7197(Iteration 8)
New Highest Accuracy: 0.72727(Iteration 10)
New Highest Accuracy: 0.73485(Iteration 37)
New Highest Accuracy: 0.75(Iteration 129)
New Highest Accuracy: 0.75758(Iteration 176)
New Highest Accuracy: 0.76515(Iteration 306)
New Highest Accuracy: 0.77273(Iteration 349)
New Highest Accuracy: 0.80303(Iteration 393)
============== Linear SVM Classifier Results ==============
Training Sample Size: (66 HGG, 66 LGG)
Testing Sample Size: (10 HGG, 10 LGG)
Average SVM Training Accuracy (1000 Iterations): 65.6364%
Highest Training Accuracy: 80.303%
Testing Accuracy with Best SVM: 70%
```

2: Sample output from SVM Classifier

Running the SVM Classifier with 1000 iterations yielded a model with a validation accuracy of 80.3% (106 out of 132 patients correctly classified). When this model was tested with the hidden dataset, the accuracy was 70% (14 out of 20 patients correctly classified).

[1] *https://machinelearningmastery.com/k-fold-cross-validation/*

Impact of Feature Selection on Accuracy

The SVM accuracy is reliant on selecting the correct radiomic features, as the differences of these features in HGG and LGG determine how well the SVM can predict future cases, using these features. If poor selectors are used, the SVM will be unable to accurately identify the difference between an LGG patient and HGG patient.

As we are using 33 different features to train our model, it is likely that some of these features will have minimal impact on the outcome of the model, and are just increasing the dimensionality of the predictor space. This may be leading to overfitting of the data. To improve on this, we could consider using Principal Component Analysis to remove the redundant features, and only keeping enough components to explain a certain percentage of variance in the model (e.g. retaining 95% variance).

Our repeatability test is calculated as follows:

1. For each acquisition protocol of a volume, the radiomic features are calculated.
2. The mean value for each feature is calculated across the acquisition protocols, and from that the average standard deviation from the mean is calculated
3. The average standard deviation is normalised across different features, to ensure comparability. We use standard deviation as a measure of repeatability, as features with low standard deviations indicates a low variance of a radiomic feature across acquisition protocols.
4. The top ten features from Intensity-, Shape-, and Texture-based radiomics are selected, by selecting the ten features with the lowest average standard deviation across the 369 volumes.

Whilst using repeatability may be a satisfactory method for binary classification, it does have some shortcomings. Using highly repeatable features is useful for selecting features that are consistent across volumes, but does not guarantee that the features are distinguishable between classifications. For example, some radiomic features, like Volume, have a high repeatability, but don't necessarily distinguish well between High-Grade and Low-Grade Glioma. This is because each MRI scan is taken at a different time of diagnosis, the time taken for the tumor to grow is not factored into the calculation.

To improve the accuracy of the SVM model, we could also consider features based on their saliency – the property that allow features of an image to be more prominent within visual clutter. Considering further criteria will allow for the selection of features that best classify the different grades.

Challenges Encountered

One issue we did encounter was deciding on which kernel function to use to train the dataset on (between Linear, Gaussian, Cubic, Quadratic, etc.). After training an SVM using each kernel function, we selected a linear kernel, as MATLAB documentation recommends it as the default for two-class classification.