An Animal Classifying ML Agent Using SVM with Orange

Group UG_6

Data Collection

Data was sourced from a publicly-available dataset* on Kaggle. This dataset contains the images of four animals common to sub-Saharan Africa: buffalo, elephants, rhinos, and zebra. Photos that are part of the dataset consist of the animals in their natural habitat and other photos that have been generated via data augmentation (e.g. photos converted to black and white). A total of 4,000 images have been sourced for the dataset, with each animal class allocated 1,000 photos.

Data Transformation / Preprocessing

Orange requires a .csv to import data, but our dataset is structured such that the names of the directories label the photos within. To circumvent this limitation, a Python script was written to generate a .csv file that obtained two points of data for each image in the dataset directory: its absolute path (**FilePath**) and the classification label associated with it (**Label**).

A **File** widget in Orange was then used to import the data. The widget was configured to select the previously-generated .csv file as input and to use the **Label** column as the **target** to indicate the variable that should be learned by the model.

The output of the **File** widget was used as input for a subsequent **Image Embedding** node. The **Image Embedding** was configured to use **Inception v3**, a pre-trained deep learning model capable of extracting sets of features from raw images.

Data Split

The **Data Sampler** widget was used to split the dataset into 70% training data and 30% test data. This widget was then used again to divide the test data into a 50-50 split of test data and validation data. To ensure replicability, the **Replicable (deterministic sampling)** option was set in both **Data Sampler** widgets. Additionally, **Stratify sample (when possible)** was enabled to ensure the equal balance between the four classes was maintained when the data was split.

Model Tuning

For the first test with test and validation data, the **Cost (C)** of the model was set to 1.00 and the **RBF** kernel was selected. In addition, the optimization parameters **Numerical tolerance** and

* https://www.kaggle.com/datasets/ayushv322/animal-classification

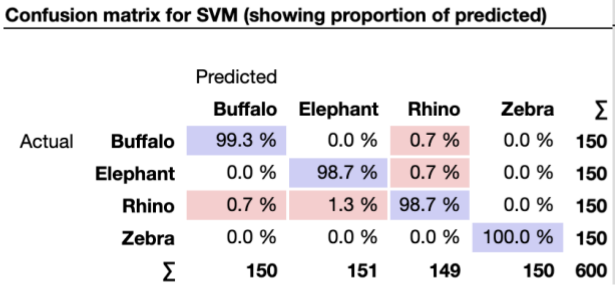
Iteration limit were set to 0.0001 and 100 respectively. The other trials involved adjusting the **Numerical tolerance**, **Iteration limit**, and **C** higher or lower. It turns out that increasing those values produced higher accuracy scores. Upon trying various combinations of values, such as **C** set to 2.00, **Numerical tolerance** to 1.00, and **Iteration limit** to 1,000 produced a higher accuracy score testing on both validation and test data by negligible amount.

Performance Evaluation

The results for the first trial are shown below in the score table from the **Test and Score** widget and the confusion matrix from the **Confusion Matrix** widget.

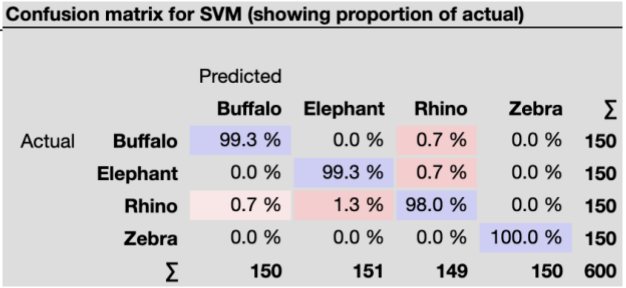
Testing the model on the test data:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	1.000	0.992	0.992	0.992	0.992	0.989



Confusion matrix for SVM (showing proportion of predicted)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	Σ
Actual	Buffalo	99.3 %	0.0 %	0.7 %	0.0 %	150
	Elephant	0.0 %	98.7 %	0.7 %	0.0 %	150
	Rhino	0.7 %	1.3 %	98.7 %	0.0 %	150
	Zebra	0.0 %	0.0 %	0.0 %	100.0 %	150
	Σ	150	151	149	150	600

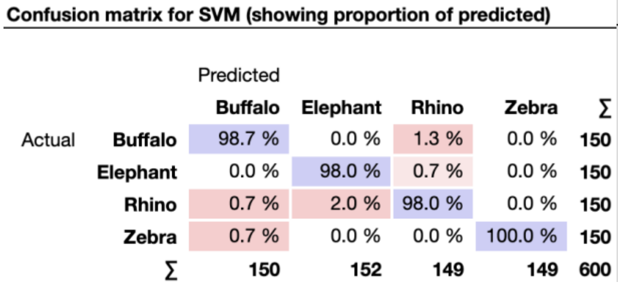


Confusion matrix for SVM (showing proportion of actual)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	Σ
Actual	Buffalo	99.3 %	0.0 %	0.7 %	0.0 %	150
	Elephant	0.0 %	99.3 %	0.7 %	0.0 %	150
	Rhino	0.7 %	1.3 %	98.0 %	0.0 %	150
	Zebra	0.0 %	0.0 %	0.0 %	100.0 %	150
	Σ	150	151	149	150	600

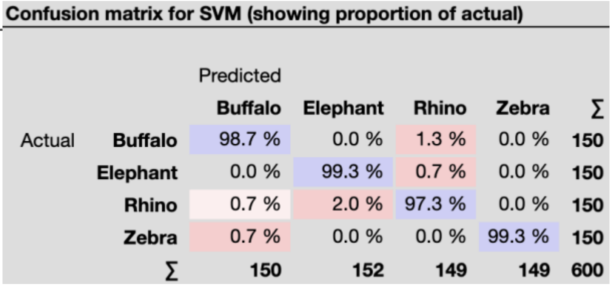
Testing the model on the validation data:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.999	0.987	0.987	0.987	0.987	0.982



Confusion matrix for SVM (showing proportion of predicted)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	Σ
Actual	Buffalo	98.7 %	0.0 %	1.3 %	0.0 %	150
	Elephant	0.0 %	98.0 %	0.7 %	0.0 %	150
	Rhino	0.7 %	2.0 %	98.0 %	0.0 %	150
	Zebra	0.7 %	0.0 %	0.0 %	100.0 %	150
	Σ	150	152	149	149	600



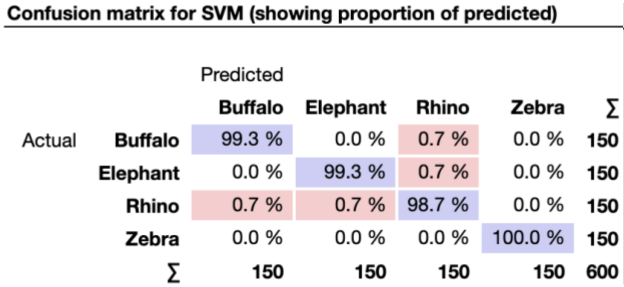
Confusion matrix for SVM (showing proportion of actual)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	Σ
Actual	Buffalo	98.7 %	0.0 %	1.3 %	0.0 %	150
	Elephant	0.0 %	99.3 %	0.7 %	0.0 %	150
	Rhino	0.7 %	2.0 %	97.3 %	0.0 %	150
	Zebra	0.7 %	0.0 %	0.0 %	99.3 %	150
	Σ	150	152	149	149	600

The results for the best accuracy score from one of the other trials are shown below in the score table from the **Test and Score** widget and the confusion matrix from the **Confusion Matrix** widget.

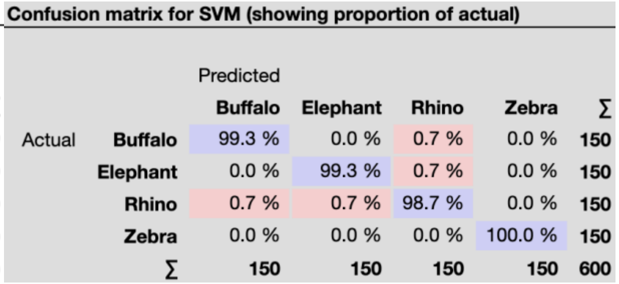
Testing the model on test data:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	1.000	0.993	0.993	0.993	0.993	0.991



Confusion matrix for SVM (showing proportion of predicted)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	∑
Actual	Buffalo	99.3 %	0.0 %	0.7 %	0.0 %	150
	Elephant	0.0 %	99.3 %	0.7 %	0.0 %	150
	Rhino	0.7 %	0.7 %	98.7 %	0.0 %	150
	Zebra	0.0 %	0.0 %	0.0 %	100.0 %	150
	∑	150	150	150	150	600

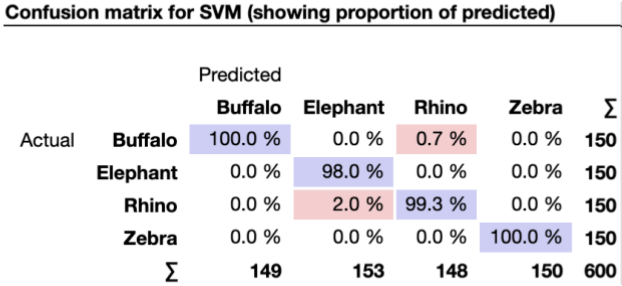


Confusion matrix for SVM (showing proportion of actual)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	∑
Actual	Buffalo	99.3 %	0.0 %	0.7 %	0.0 %	150
	Elephant	0.0 %	99.3 %	0.7 %	0.0 %	150
	Rhino	0.7 %	0.7 %	98.7 %	0.0 %	150
	Zebra	0.0 %	0.0 %	0.0 %	100.0 %	150
	∑	150	150	150	150	600

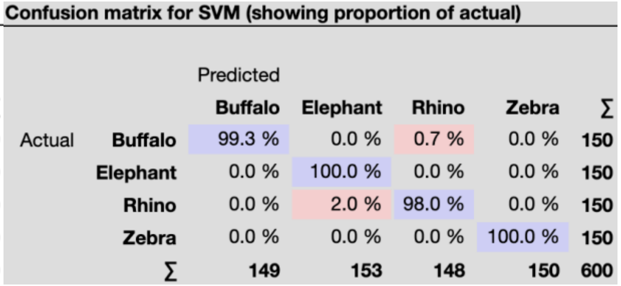
Testing the model on validation data:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.999	0.993	0.993	0.993	0.993	0.991



Confusion matrix for SVM (showing proportion of predicted)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	∑
Actual	Buffalo	100.0 %	0.0 %	0.7 %	0.0 %	150
	Elephant	0.0 %	98.0 %	0.0 %	0.0 %	150
	Rhino	0.0 %	2.0 %	99.3 %	0.0 %	150
	Zebra	0.0 %	0.0 %	0.0 %	100.0 %	150
	∑	149	153	148	150	600



Confusion matrix for SVM (showing proportion of actual)

		Predicted				
		Buffalo	Elephant	Rhino	Zebra	∑
Actual	Buffalo	99.3 %	0.0 %	0.7 %	0.0 %	150
	Elephant	0.0 %	100.0 %	0.0 %	0.0 %	150
	Rhino	0.0 %	2.0 %	98.0 %	0.0 %	150
	Zebra	0.0 %	0.0 %	0.0 %	100.0 %	150
	∑	149	153	148	150	600

Contributions

Carlos Escudero: Evaluated various publicly-available datasets on Kaggle and contributed edits and content to the project report.
Hunter Evans: Assisted in locating a viable dataset for the project. Created and managed the GitLab repository used as part of the project. Contributed edits to report.
Brian Lee: Created the final SVM using Orange and wrote the initial report draft.
Drew Wheeler: Prototyped SVM in Orange and created revision of initial report draft.
Robert Zheng: Provided feedback for chosen datasets. Created slidedeck used for presentation.