Homework #5: Image Analytics

Jana Obšteter, **jana.obsteter@kis.si**

Data: Data included 69 images displaying three types of cells – animal, plant and fungal. The three cells have different characteristic: animal cells have a large nucleus and no cell wall, plant cells have cell walls, vacuoles and chlorophyll and are tightly packed; and fungi cell are organised in a chain-like shape called hyphae. Each class was represented with 23 images.
The images were first cropped to 299x299 size and embedded using inception v3.

Methods

1.Clustering

First we used **hierarchical clustering** to classify the types of cells. The resulting dendrogram is presented in Figure 1. We determined the three-clusters threshold on the dendrogram and inspected the results. One of the clusters (green) consisted from 21 plant cell images and one animal. Another cluster of 15 images (red) consisted almost only from animal cells images, with the exception of one fungal and one plant cell image. The third cluster (blue) consisted of 22 fungal cell images with additional four animal and one plant cell image.

The results show that classes of cell images have characteristics that are unique enough for the hierarchical clustering to be able to differentiate between them. Apparently the characteristics of fungal and plant cells are more distinct than those of animal cells, since only one fungal and two plant cells were misclassified compared to the five misclassified animal cells. When inspecting the plant and fungal images that were misclassified as animal cell images, we see a similarity in the shape and size of the cells – all the images in this cluster have a large number of small cells.

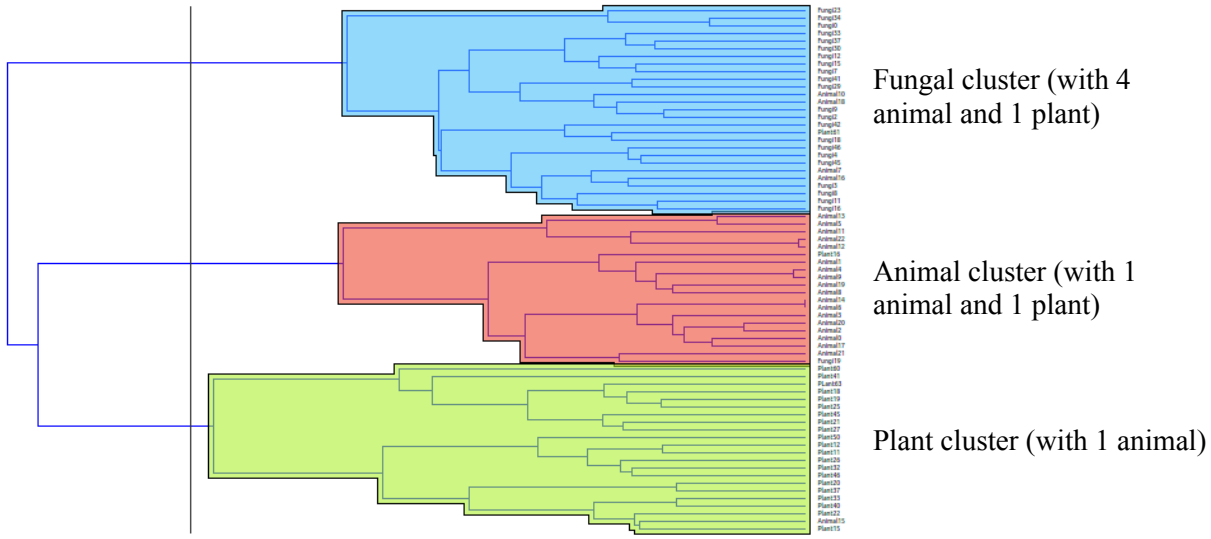

Figure 1: Dendrogram of the hierarchical clustering of images displaying animal, plant and fungal cells.

We also performed **k-means** on the data. It performs slightly worse that the hierarchical clustering. Here the most misclassified images are those of fungal cells and the less those of plant cells.

2. Learning algorithms

Next, **learning algorithms** were applied to the subclasses of images and the success was inspected with 3-fold cross validation. The 3-fold cross-validation was used due to a smaller dataset in order to reduve the predictoin variance. If we used 10x CV, only two images would be left per class for predicting the class – therefore it could be only 0, 50% or 100% and the prediction variance would be large. Also, the chance of

assigning the true class by chance is high. The following algorithms were applied: **logistic regression, kNN, SVM and random forest.**

The learning algorithms allow for a multi-class classification and therefore can be used for classifying the images based on their embedded features. The only exception are trees, which fail to classify the data since they are built for binary classification. But even this can be overcome - when repeating the process of a binary split, such as in **random forest** with many trees, each time a different set of the two (out of three) classes are predicted. So if we repeat the process enough times, even the **random forest** performs well on classifying the three classes. The quality parameters are still lower than with the other learners – AUC = 0.906, precision = 0.818 and recall 0.783. This approach would however not work for a larger number of classes since this would require a much larger number of trees in the random forest.

The **kNN** performs poor with Euclidean distance. However, by tweaking the parameters and choosing Manhattan distance with 10 nearest neighbors, the kNN reaches the AUC of 0.979. The precision is 0.875 and recall is 0.913. Since the recall is slightly higher, this means that the number of false positives is a bit higher than the number of false negatives. There is only one misclassified animal cell image, two fungal and four animal cell images. The Manhattan distance measures the distance between two points in a grid based on a strictly horizontal and/or vertical path. This could be useful if the input variables are of different type. We also chose distance instead of uniform which weights neighboring points according to their distance instead of giving them all equal importance.

SVM performs best when choosing polynomial function with degree of two (after two the quality parameter decrease). This means that data classes can be separated by transforming the data into a quadratic space. The AUC of this SVM is 0.977 with precision of 0.880 and recall 0.957. Again, most of the images are in the right classes with a couple of additional images in each class. There are two plant and one fungal misclassified images, both misclassified by kNN as well. There are four misclassified animal cell images, not all being the same as in kNN.

Logistic regression is also relatively successful in classifying the images. The Ridge normalisation performs slightly better than the Lasso regression. The AUC for the Ridge regression is 0.978 with precision and recall of 0.78 and 0.87, respectively. Logistic regression misclassified two plant cell, two fungal and four animal cell images. One of each – animal, plant and fungi cell image – was erroneously predicted by kNN and SVM as well..

In general, by tweaking the parameters of the learners, we can reach acceptable accuracy and AUC for all the learning algorithms. For all of them the recall is higher than the precision. When inspecting the missclasified cases between the learners we observe there is a large overlap – i.e. all the learners misclassify the same images. If we look at these images we can – even with a naked eye – see that they resemble more that of another class. For example, most images of plant cells display larger cells while that of animal display smaller cells. Therefore a image displaying small plant cells was misclassified as animal cell image.

3. Projection into 2D space

PCA: PC1 explains 8% of the data variance and PC2 another 6%. By inspecting the plot of PC1 against PC2 we see that the first two principal components explain enough variance for us to be able to roughly identify the classes. The most clearly defined is the "Plant" class and this is in concordance with the results of learning algorithms (plant cell images were the least misclassified). In the middle there is and overlap of the classes. By inspecting these data points we see that there are the same images that were misclassified by logistic regression, kNN and SVM (circled points).

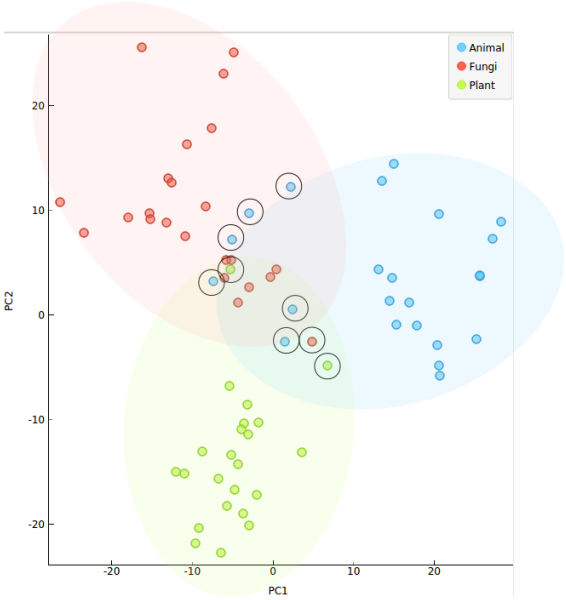

Figure 2: PCA on the image embedding data of animal, plant and fungal cell images.

4. Linear regression and trees

Tree can not perform multiclass classification based on many numerical outputs and a single class output. We also can not perform linear regression since the dependent variable is a class variable. By creating a set of numerical (continuous) features that determine (numerically describe) the classes we can perform linear regression for classification or use trees to predict the classes. The quality of such classification is measured by a mean squared error (MSE) and R^2 of the model.

In this dataset we have created three numerical target variables – Cell Wall, Vacuoles and and Hyphae. Value of 1 indicates the presence of the feature in the cell type. Value of 1 for cell wall was assigned to plant and fungal class, 1 for vacuoles to plant class and 1 for hyphae to fungal class. All other values were 0. Again, although these features are binary, they were coded as numerical for the purpose of linear regression. The head of the table is shown in Table 1. The explanatory variables included the features from the images embedding.

Table 1: The head of the table with additional features for linear regression. Hyphae, CellWall and Vacuoles represent created numerical features and n1 - nx represent the features from the image embedding.

Category	Hyphae	CellWall	Vacuoles	n1	n2	n3	n4
d	d	d	d				
class							
Plant	0	1	1	36341	441	398	0
Plant	0	1	1	23619	299	299	0
Plant	0	1	1	42634	400	300	5
Plant	0	1	1	22849	299	299	30

We could run **linear regression** and **tree classification** on each one of the created numerical features. By doing this we get R2 of the models for prediction the target features – that is the presence of vacuoles, cell wall or the hyphaes. This regression does not allow to distinguish between the three classes simultaneously, but only between the classes that do / do not posses the features. For examples, by regression on hyphae (1 / 0) we can distibguish between fungal and other (animal and plant) cells. The R2 are shown in Table 2. We see that linear regression performs better than the classification tree which means that vertical and horizontal splits do not split the data well. The best split of the data is on the cell wall features that splits animal from plant and fungal class.

Table 2: The coefficients of determination (R2) for the linear regression and classification tree on the created numerical target features.

	Vacuole	Cell Wall	Hyphae
Linear regression	0.782	0.795	0.651
Tree	0.266	0.446	-0.255

By combining the results of the three regression we could determine the class of each image.

Conclusion

Extracting image features through image embedding provides us with data to classify the images. The latter could be done with either clustering, projection into 2D space or learning algorithms. Tweaking the parameters of the algorithms allows us to find the optimum method which should be determined through cross validation or by looking at some other quality parameters.