CSC2005Z Research Proposal

Luc Hayward – HYWLUC001

Classification of tree presence in a region at a per pixel level

A comparison of the effectiveness of Support Vector Machines and Random Forest classifiers for tree detection

1. Project Description

Aerobotics are an agricultural start-up providing early disease detection and crop monitoring solutions to farmers through a combination of satellite imagery and high-resolution drone scans to identify the size and location of trees for agricultural monitoring. (Aerobotics, 2018)
In addition to high resolution RGB images, each survey also includes near infrared (NIR) images as well as height maps of the area. The NIR data is used to produce a normalized difference vegetation index (NDVI), an estimate of photosynthetic absorption and thus vegetation density. (Raj & SivaSathya, 2014)

Currently Aerobotics has been required to largely mask out the locations of trees manually for each of their surveys. However, this has led to an interest in the potential for machine learning to aid in the process of masking the location of trees in the image with their previous scans available for use as training samples.

Image classification is an ongoing area of research relating to computer vision. Due to the large range of different problems it can be applied to and the trade-offs between different approaches, there is not any one method which is conclusively better than all others. (Sonka, et al., 2014)

By using machine learning, it is hoped that an accurate classifier can be created which will be able to match the accuracy of the current masks produced by Aerobotics without increasing time taken to create the mask. This project aims to investigate and compare the effectiveness of two different classes of classifiers, SVM and Random forests, to reduce the reliance on manual masking techniques currently in use. SVM and Random Forest classifiers are widely used for classification and give good performance across a variety of tasks including image classification. (Raczko & Zagajewski, 2017) By incorporating NDVI data, it is hoped that accuracy can be improved beyond simply using the RGB data for the classifiers.

2. Research Questions

Which style of classifier, SVM or Random Forests, is best able to mask out the location of trees in a scan.

Current approaches for classifying the location of trees in a given region has found Aerobotics resorting to masking out their scans by hand. It is hoped that by applying a machine learning (ML) classifier, in this case either SVM or Random Forest, a more automated classification method can be produced reducing the reliance on manual masking techniques.

Ultimately, a good ML implementation would produce a mask of the scans indicating the regions in which trees were present to a degree of accuracy as similar as possible to current results. In

order to produce such a mask, the classifier would need to be able to process a given scan and produce a binary result of tree present or not present on a per pixel basis.

A subset of the data provided by Aerobotics can be used to validate the effectiveness of the classifiers. As the masks used for validation are not necessarily one hundred percent perfect, the absolute accuracy of the classifiers will depend upon the quality of the ground truth masks chosen.

A successful classifier would be testable against the most accurate masks provided by Aerobotics and achieve an 80% similarity in the masks produced using the Intersection over Union (IoU) method. (Rahman & Wang, 2016) Using this test, as both classifiers will be trained on the same data, the most effective model should consistently produce better results for the given problem.

3. Methodology

Background research will be conducted to establish the necessary algorithms, frameworks and tools necessary for this project. A number of scans will be provided by Aerobotics, including hand drawn masks of where trees are in each scan. Unfortunately, as they have been done by hand, some of the masks are too inaccurate to be used without throwing off the training or results. The scans will need to be separated into two sets with those containing the most accurate masks used for validation and training and the unusable data set aside.

The classifiers will be implemented and trained on a subset of the most accurate data with a small number of scans kept in reserve to be used for final testing of the models. During the validation phase, K-fold cross validation will be used to ensure the model remains applicable to new data. This style of cross validation will be used as it is commonly chosen and generally results in a less biased or "optimistic" estimate of the model's accuracy as well as accounting for variance in the results.

The classifiers will be tested using both intersection of union and a confusion matrix (HAY, 1988) to determine their advantages and disadvantages. Different hyper-parameters of the models can be tweaked to try to optimise the different classifiers after initial testing and improve performance. Finally, a report will be drafted summarising the results and discussing the proposed optimal solution

4. Work Detail

4.1. Risks

Risk	Probability	Impact	Factor	Mitigation Strategy
Data loss	1	10	10	Multiple backups including offsite storage for data and source code
Lack of suitable software	7	8	56	Custom software development
Classifiers unsuitable for given dataset	5	7	35	Research classifiers in advance, explore reasons classifiers might have failed.

4.2. Timeline

Month	August					September				October				November	
Day	30-5	6-12	13-19	20-26	27-2	3-9	10-16	17-23	24-30	1-7	8-14	15-21	22-28	29-4	5-11
Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Task															
Project Proposal															
Background Research															
Software Installation and Preparation															
Data exploration															
Classifier implementation															
Classifier testing															
Classifier comparisons															
Find optimal solution															
Research Report – Rough Draft															
Final Research Report															

4.3. Resources Required

- Python
  - SciKit Learn (scikit-learn, 2018)
  - NumPy (NumPy, 2018)
  - SciPy (SciPy, 2018)
  - Matplotlib (Matplotlib, 2018)
- QGIS (QGIS, 2018)
- Storage for scan data

4.4. Deliverables

- Research proposal
- Research report

4.5. Milestones

1. Topic chosen
2. Classifier styles chosen
3. Software environment preparation complete
4. Separate data into training data and testing data
5. Present research proposal
6. Working classifiers
7. Add more features into the classifiers
8. Multiple classifiers with different parameters
9. Optimal parameters set
10. All coding completed
11. Comparison of classifiers
12. Optimal classifier chosen
13. Research Report Rough Draft
14. Final Research Report

4.6. Ethical Issues

There are no major or obvious ethical issues in this case. The data does not contain any identifiable information and scans do no contain information linking them to specific farms or companies.

5. Evaluation of Research Questions

In order to compare the performance of the classifiers, intersection of union tests and confusion matrixes will be calculated for each classifier. Using k-fold cross validation in conjunction with these methods, the possibility of incorrectly estimating bias and variance in the results is reduced.

Intersection of Union provides an empirical value representing classifier accuracy as the overlap between two different masks. Higher IoU values indicate that the classifier produced a mask where a greater portion of the predicted area containing a tree lined up with the ground truth area. This would indicate that the classifier had correctly learned to determine whether a tree was present in a region on a per pixel basis.

Additionally, a “confusion matrix” will be tabled allowing the calculation of true positives and negatives and false positive and negatives, further indicating the effectiveness of each classifier. This will allow for metrics such as accuracy, specificity and precision to be calculated.
This is similar to IoU but is specifically designed for use in the testing of binary classifiers as it essentially looks at the percentage of pixels correctly classified.

By testing against a separate set of data after validation, it can be determined whether the classifiers were able to learn a general model without overfitting the training data and which of the two classifiers is best able to be applied to new data.

6. Anticipated Outcomes

After all implementations and testing of classifiers is completed, the following final outcomes are expected:

- A classifier that meets or exceeds an Intersection of Union of 0.8
- It is predicted that the Random Forests approach will be most successful
- Visual comparison of masks produced

Bibliography and Previous Systems

Works Cited

Aerobotics, 2018. *Aerobotics.* [Online]
Available at: https://www.aerobotics.io
[Accessed 31 August 2018].

Duro, D. C., Frnaklin, S. E. & Dubé, M. G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment,* 15 March, Volume 118, pp. 259-272.

HAY, A., 1988. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing,* 9(8), pp. 1395-1398.

Matplotlib, 2018. *matplotlib.* [Online]
Available at: www.matplotlib.org
[Accessed 04 09 2018].

NumPy, 2018. *NumPy.* [Online]
Available at: www.numpy.org
[Accessed 04 09 2018].

QGIS, 2018. *QGIS A Free and Open Source Geographic Information System.* [Online]
Available at: https://qgis.org/en/site/
[Accessed 04 09 2018].

Raczko, E. & Zagajewski, B., 2017. Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sending,* 50(1), pp. 144-154.

Rahman, A. M. & Yang, W., 2016. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation.. In: *Lecture Notes in Computer Science.* s.l.:Springer, Cham.

Raj, J. K. & SivaSathya, S., 2014. SVM and Random Forest Classification of Satellite Image with NDVI as an Additional Attribute to the Dataset. *Proceedings of the Third International Conference on Soft Computing for Problem Solving,* Volume 258, pp. 95-107.

scikit-learn, 2018. *scikit-learn.* [Online]
Available at: http://scikit-learn.org/stable/index.html#
[Accessed 04 09 2018].

SciPy, 2018. *SciPy.org.* [Online]
Available at: www.scipy.org
[Accessed 04 09 2018].

Sonka, M., Hlavac, V. & Boyle, R., 2014. *Image processing, Analysis, and machine learning.* 4th ed. s.l.:Cengage :earning.