RMIT at ImageCLEF 2011 Plant Identification

Rahayu A. Hamid and James A. Thom

School of Computer Science and Information Technology,
RMIT University, Melbourne, Australia
rahayu.ahamid@student.rmit.edu.au,james.thom@rmit.edu.au
http://www.rmit.edu.au

Abstract. This paper presents the contribution of the ISAR group at RMIT University to the ImageCLEF 2011 Plant identification task. The task involves identifying various different species of trees based on images of their leaves. Our main objective is to investigate the performance of two classification algorithms in associating the correct tree species to each test image. We extracted visual features from the data set using the feature extraction module in GIFT. From all the features extracted, we selected 166 features of the colour histogram. The classification algorithms used are instance based learning and decision trees. Both algorithms were implemented using the Weka 3 data mining toolkit. Classifiers for both algorithms were evaluated by a 10 folds cross-validation. Based on the official results, our runs did not perform well due to three main reasons namely, feature selection, training data and classifier parameters.

Keywords: Plant identification, Image feature extraction, Classification

1 Introduction

This paper presents the participation of the ISAR group (Information Storage Analysis and Retrieval) at RMIT University in the ImageCLEF 2011 Plant Identification task. The task was motivated by the need to accurately gather knowledge of the identity, geographic distribution and uses of plants in ensuring advancement in agriculture and safeguarding its diversity.

The main goal of the task is to correctly associate tree species to each test image. The task is treated as a supervised classification problem with tree species used as class labels. Our objective in the task, however, is to investigate the performance of two classification algorithms in classifying the test images to the tree species.

The pilot task dataset contains approximately 5400 pictures of leaves from 71 tree species from French Mediterranean area. Further details regarding the general setup of the dataset are available in the task description [1]. The rest of this paper is organised as follows: Section 2 describes the experiment carried out, Section 3 the results we obtained at ImageCLEF 2011, then we conclude the paper.

2 Experiments Description

2.1 Feature Extraction

Classification can be done by either using textual features (from the XML files), visual features (from the jpg files) or combination of both textual and visual features. Our work is based on visual features only. We extracted the visual feature from the data set using the feature extraction module in the GNU Image-Finding Tool (GIFT) [2]. The total number of features extracted by GIFT is approximately 80,000 features. From these features, we selected only 166 colour histogram features. GIFT uses a palette of 166 colours derived by quantising the HSV colour space into 18 hues, 3 saturations, 3 values and 4 grey levels [3]. Histogram intersection is used to measure the distance/similarity between colour in the images. Colour histogram was chosen because each image has its own colour distribution in the colour histogram, which will be able to distinguish it from other images. Furthermore, as we are experimenting with basic classification algorithms, using colour histogram features seems reasonable [4].

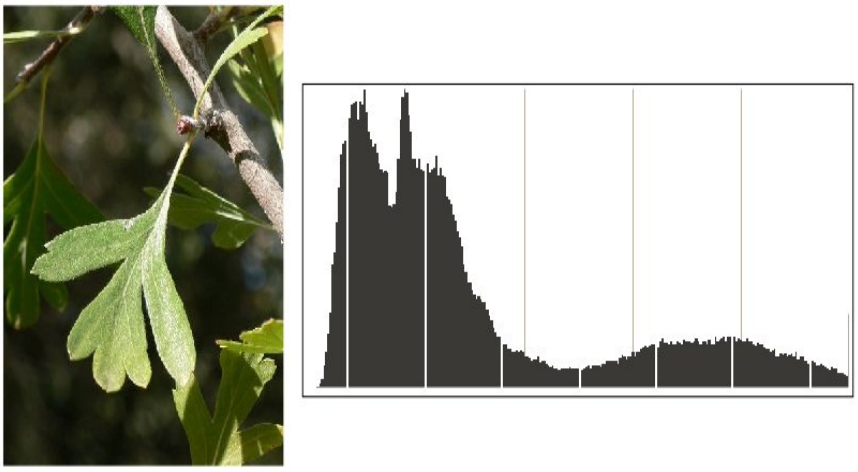

Fig. 1. Example of a free natural photo of a tree with its colour histogram

2.2 Classification Algorithms

Our classification algorithms were implemented using the Weka 3 data mining toolkit [5]. Prior to deciding which classification algorithms to use, we trained several different classification algorithms that are available in Weka. The purpose of this acticity is to identify classifier(s) that produces the highest classification rate. Five types of classifier were trained, namely Bayesian, decision tree, instance-based learning, rules and functions.

The classifiers were trained using all the training data together, without separating them according to the type of image acquisition. In order to reduce variability and estimate how accurately the classifier will perform, they were evaluated by a 10-folds cross validation. Note that Weka support several instance based algorithms namely IB1 and IBk whereby k = *2,....,n*. Table 1 shows the classification rate for the different classifiers trained.

Table 1. Classification rate of different types of classifier on a 10-folds cross validation

Type	Classifier	Classification rate
Bayesian	NaiveBayes	33.03%
Decision tree	J48	53.20%
Instance based learning	IB1	60.01%
	IBk (k=2)	55.61%
Rules	JRip	46.95%
Functions	SMO	34.51%

From the table, we can see that decision tree and instance based learning classifier performs better than the rest. Although IBk performed slightly better than J48, we selected IB1 and J48 so as to compare between the two different classifiers.

The IB1 algorithm is identical to the nearest neighbours algorithm. It is considered as a statistical learning algorithm and is simple to implement. When asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training-point accordingly to some distance metric [6].

A decision tree partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label, a value or an action to characterise its data points. It is used to classify a case by starting at the root of the tree and moving through it until a leaf is encountered. At each non-leaf decision node, the case's outcome for the test at the node is determined and attention shifts to the root of the sub-tree corresponding to this outcome. When this process finally leads to a leaf, the class of the case is predicted to be that recorded at the leaf.

The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. However, many decision tree construction algorithms involve a two-step process. First, a very large decision tree is grown. Then, to reduce its large size and over-fitting the data, in the second step, the given tree is pruned [7]. The pruned decision tree that is used for classification purposes is called the classification tree.

The Weka 3 implementation of IB1 classifier uses normalised Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as this training instance. If multiple instances have the same distance (closest) to the test instance, the first instance found is used. The difference between IB1 and IBk is that there are no parameters that could be changed.

As for the decision tree classifier, J48 is Weka's implementation of the C4.5 algorithm. The C4.5 decision tree can be either a pruned or unpruned tree. In our experiment, we created a pruned decision tree. We did not use binary splits when building the trees. The confidence factor used for pruning the tree was 0.25 with the minimum number of instance per leaf set as 2. In determining the amount of data used for pruning (number of folds), we used the default value 3. We considered the subtree raising operation when pruning and did not smooth the object counts at the leaves.

3 Results

The objective of our experiment is to evaluate the effectiveness of both our classifiers in classifying tree species based on images of its leaves. We submitted two runs, one for each classifier. RMIT_run1 used the instance based learning IB1 classifier while RMIT_run2 used the decision tree classifier J48. As shown in Table 2, our first run, RMIT_run1 performed slightly better in terms of average images identified for each type of image acquisition. However, it was unable to identify images of the scan-like type.

Table 2. Results of our submitted runs for the Plant identification task

Runs	Scan	Scan like	Photograph	Mean
RMIT_run1	0.071	0.000	0.098	0.056
RMIT_run2	0.061	0.032	0.043	0.045

The overall results of participating groups that had submitted runs for the task are in the task description [1].

4 Conclusion

Our group submitted two runs in our first participation in the ImageCLEF 2011 Plant identification task. The results obtained by our runs were poor. This is due to three main reasons. The first is poor selection of features. We only used visual features which is the colour histogram and it was not suitable in identifying images based on the type of image acquisition used in the task. Next, we used all the training data together to train the classifier instead of dividing them according to the type of image acquisition. Finally, we did not exhaust all the parameters used in training both of the classifiers. We hope to further improve our experiment in future tasks.

References

1. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthelemy, D., Molino, J.-F., Birnbaum, P., Mouysset, E., Picard, M.: The CLEF 2011 plant image classification task. CLEF 2011 working notes. Amsterdam, The Netherlands, (2011)

2. GIFT : The GNU Image-Finding Tool, http://www.gnu.org/s/gift/
3. Squire, D.M., Müller, W., Müller, H., Thierry P.: Content-based query of image databases: inspirations from text retrieval. Pattern Recognition Letters. 21, 1193–1198 (2000)
4. Deselaers, T., Keysers, D., Ney, H. : Features for image retrieval: an experimental comparison. Information Retrieval. 11, 77–107 (2008)
5. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
6. Aha, D.W., Kibler, D., Albert, M.K. : Instance-based learning algorithms. Machine Learning. 6, 37–66 (1991)
7. Quinlan, R. : C4.5: Programs for Machine Learning. Morgan Kaufmann , San Mateo, CA (1993)