Hierarchical Classification of Sea-Floor Imagery

M.S. Bewley[1], N. Nourani-Vatani[1], B. Douillard[2], O. Pizarro[1], S.B. Williams[1]
[1]Australian Centre for Field Robotics
The University of Sydney, NSW, Australia
m.bewley@acfr.usyd.edu.au

[2]Jet Propulsion Laboratory, CA

1. Background

Supervised classification on data sets with hierarchical labels raises a number of challenges not present in binary or flat multi-class classification problems. An example of this type of problem is in scientific imaging of the sea-floor. To study ecosystems and populations, marine scientists require quantitative data on bottom-dwelling organisms and physical morphology. The state of the art is to take a selection of images, manually label the content, and extrapolate to assess distribution and coverage over wider geographical areas.

The image content is labelled in a hierarchical manner; a standardised tree structure for biological and physical classes was recently defined in the Catami project [2]. Typically 50 randomly located keypoints in an image will be labelled, and the class assigned may be at a higher level (such as "unspecified biology"), all the way to leaf node classification of particular species.

Recent work on a simplified attempt at automated labelling is described in [1]. The data set was from an Autonomous Underwater Vehicle (AUV) campaign in Tasmania (Australia) in 2008. The manual labelling effort resulted in 62,900 hand labelled points from 1,258 geographically dispersed images (with many more images remaining unlabeled). We use the same data set in this study, but extend the problem from single species classification (Kelp) to a hierarchy of up to four layers with 19 species and physical morphologies (using the Catami hierarchy, including classes such as coral, sponges, soft-physical, algae, and others). Due to the way the data was labelled (many keypoints were not identified to the deepest valid node), it was not possible to apply one-vs-all flat multi-class classification to the whole problem. Discriminating a node from its ancestor classes is not valid in this case, as the ancestors are likely to include data that truly belongs to a leaf node. This type of data can occur for a number of reasons: Limited expertise of the labeller, the specific scientific interest of the group funding the labelling, or even the ability to identify specific species from the local visual information available in the image.

Further, the nature of the data requires supervised learning, in order to meet the scientists' needs. Regardless of whether the semantic hierarchy is the most optimal for learning, or the most cleanly separated based on visual data, we need to speak the same language as the marine scientists, to ensure the output is directly useful for them. As we are required to conform to a pre-defined hierarchy, techniques constructing hierarchies in an purely unsupervised manner are not appropriate.

2. Method

Work on this type of supervised hierarchical classification problem was reviewed extensively in [5]. They defined a number of approaches; we chose to test the approach they describe as "by far the most used in the literature", called *Local Classifier per Node*. Each node in the classification tree has a binary classifier that is trained to distinguish that class from others. An important decision is the definition of positive and negative training examples for each node. We compare the two policies described in [5] that most naturally fit our problem: The *inclusive policy* includes the entire subtree of the training node as positive examples, with nodes in the rest of the tree (with the exception of direct ancestors of the training node) as negative examples. This policy represents the most data that can be validly used for training a given node's classifier. The alternative *sibling policy* uses the same positive training examples, however the negative examples are restricted to siblings of the training node (and not siblings of the ancestor nodes). The expected performance difference between these two policies is not obvious, with no clear winner found in [3]. On one hand, the *inclusive policy* ensures that each node is as informed as possible, and should be able to deal better with classifying instances that belong elsewhere in the tree. On the other hand, the *siblings policy* solves a much more specific problem (distinguishing a node's class from its sib-

lings), and may give better discriminative performance between these classes. Such nodes will, however, be less informed about instances that belong elsewhere in the tree. An inherent advantage of the latter approach is that far less training data is required for nodes deeper in the tree, which becomes significant when the tree is large.

A subtlety of the *siblings* vs. *inclusive* approaches is in the selection of image features. In flat multi-class classification, the same image features are typically used for all classifiers. With the hierarchical class structure, we can select (either manually, or by feature learning techniques) features that are optimised for the siblings or inclusive training data sets.

After deciding on the classifier structure and training data sets, we must still choose a technique for predicting instance classes. If we require complete consistency in the hierarchical labels (such that there is a single, unbroken chain of classifiers predicting positive results from the root to the deepest node), the simplest choice described in [5] is what we refer to as *max probability switching (MPS)*. An instance starts at the root node, and flows to the child node with the highest prediction probability (akin to performing *one-vs-rest* classification at each node). This technique implies prediction down to the leaf node level. We can remove this constraint by stopping an instance from moving further down the tree when the maximum predicted probability falls below some threshold (say 0.5 for even weighting). We also test an alternative approach - the use of a simple probabilistic graphical model (PGM), where the class tree also represents the independence relations in the PGM, and we assume the conditional probability of node membership is given by the probabilistic predictions of the classifiers. This allows exact inference to be trivially performed, by multiplying probabilities of a leaf node's ancestors to obtain the probability of membership.

Lastly, a robust performance metric is needed; ideally a single number to evaluate the performance on an entire tree. The closest commonly used in the literature is the hierarchical f1-score [4]. Each instance has multiple counts of true/false positives/negatives, as each node in the chain of true class nodes is compared to the chain of predicted nodes. We modify this metric such that if the predicted class is more specific (lower down the tree) than the true class, we do not apply false positive penalties. Given the manner in which the data was labelled, it would be unfair to reward or penalise any results deeper than the deepest known class.

We present results using logistic regression (LR) classifiers, with features derived using PCA and Local Binary Patterns (LBP) on the Tasmania 2008 data set. Performance is measured in terms of the modified hierarchical f1-score with the same training and validation sets described in [1].

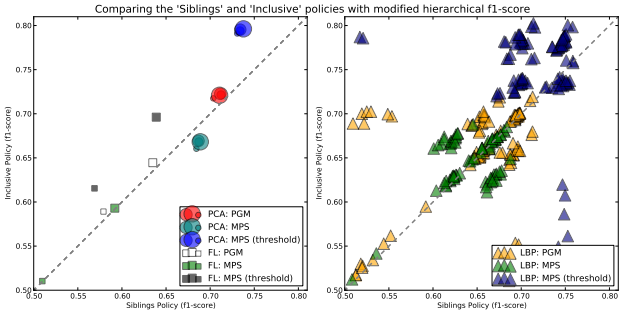

Figure 1. Comparison of *Siblings* and *Inclusive* Policies (on modified hierarchical f1-score). Each marker represents both local and global performance on a given feature and prediction setup.

3. Results and Discussion

We compare the *sibling* and *inclusive* policies on a range of image descriptors. In Figure 1, performance using the modified hierarchical f1-score is compared for the two policies, across a range of image descriptors and prediction approaches. PCA uses 60 components on each of 7, 15, 31, 63 and 95 pixel square RGB image patches. LBP uses a variety of colour space transforms and LBP types (uniform, rotation invariant, fourier) on a 31 pixel patch.

There is a clear trend that for mandatory leaf node prediction, the *PGM* is generally superior to *MPS*. This is promising for future work, as the PGM is the more principled approach, and more sophisticated models can be used. Also, the modification to permit the network to predict only higher level classes when less confident (thresholding) was highly successful.

In terms of *inclusive* and *sibling* policies, we obtain the same finding on underwater images as that found using text classification [3] — no clear winner. Given *sibling* has a significant advantage in reducing training time, it is preferred in situations where the results are comparable.

With these results, future work with PGMs with variable tree-depth prediction would be valuable. Also, it may be possible to further boost the performance of the *sibling* relative to the *inclusive* policy with the use of more sophisticated local classifier optimisations.

References

[1] M. Bewley, B. Douillard, N. Nourani-Vatani, A. Friedman, O. Pizarro, and S. Williams. Automated species detection: An experimental approach to kelp detection from sea-floor AUV images. Dec. 2012. 1, 2

[2] L. Edwards. Release of CATAMI classification scheme, Feb. 2013. 1

[3] T. Fagni and F. Sebastiani. On the selection of negative examples for hierarchical text categorization. In *Proceedings of the 3rd Language & Technology Conference (LTC'07)*, pages 24–28, 2007. 2, 3

[4] S. Kiritchenko, S. Matwin, R. Nock, and A. Famili. Learning and evaluation in the presence of class hierarchies: application to text categorization. *Advances in Artificial Intelligence*, pages 395–406, 2006. 2

[5] C. N. Silla Jr and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011. 2