Preprint from *SPIE Defense + Commercial Sensing, Signal Processing, Sensor/Information Fusion, and Target Recognition XXXI*, 3-7 April 2022.

What is the Best Level to Report from a Hierarchical Classifier?

Mark Carlotto (mark@carlotto.us)

ABSTRACT

In a multi-level (coarse to fine) classifier, decisions along with their estimated uncertainties can be reported at any or all levels. We define the concept of the “best” level from the standpoint of a user who seeks to make a single class call that seeks to maximize specificity and minimize the error in making a mistake.

Introduction

Hierarchical classifiers define a set of classes organized in a coarse-to-fine manner in trees, i.e., directed acyclic graph where children have only one parent. Chavez (2018) notes that breaking down a classification problem in this way can increase classification performance at the expense of system complexity. It also adds to the complexity in which the results can be reported. For example, should the most likely class at the finest (most detailed) level be reported, or the most likely class at all levels? Reporting the most likely class at the finest level may not be the best choice in circumstances where the likelihood of its parent class is not significantly higher than another class at that level. Reporting results at all levels may provide too much information.

Wu and Tygert (2019) describe metrics of failure (so-called "loss" or "win") that penalize classes based on similarity; e.g., it is better to misclassify a sheepdog as a poodle than a skyscraper. As one descends a tree, classes become more specific (child classes under the same parent become more similar to one another). We introduce the concept of reporting at the “best” level in a hierarchy that is intended to provide a single common-sense result for users that lack analytical skills or simply seek a simple answer. Finding the best level is formulated as the problem of finding the minimum value of an objective function that considers the specificity of a decision and the loss or probability that it could be the wrong decision.

Approach

We define the label space to be a subdivision of the classification space into a tree of labels, label support to be the “area” of a label (coarser labels are more general and

have greater support than finer labels), and label probabilities to be the computed class likelihoods (Figure 1).

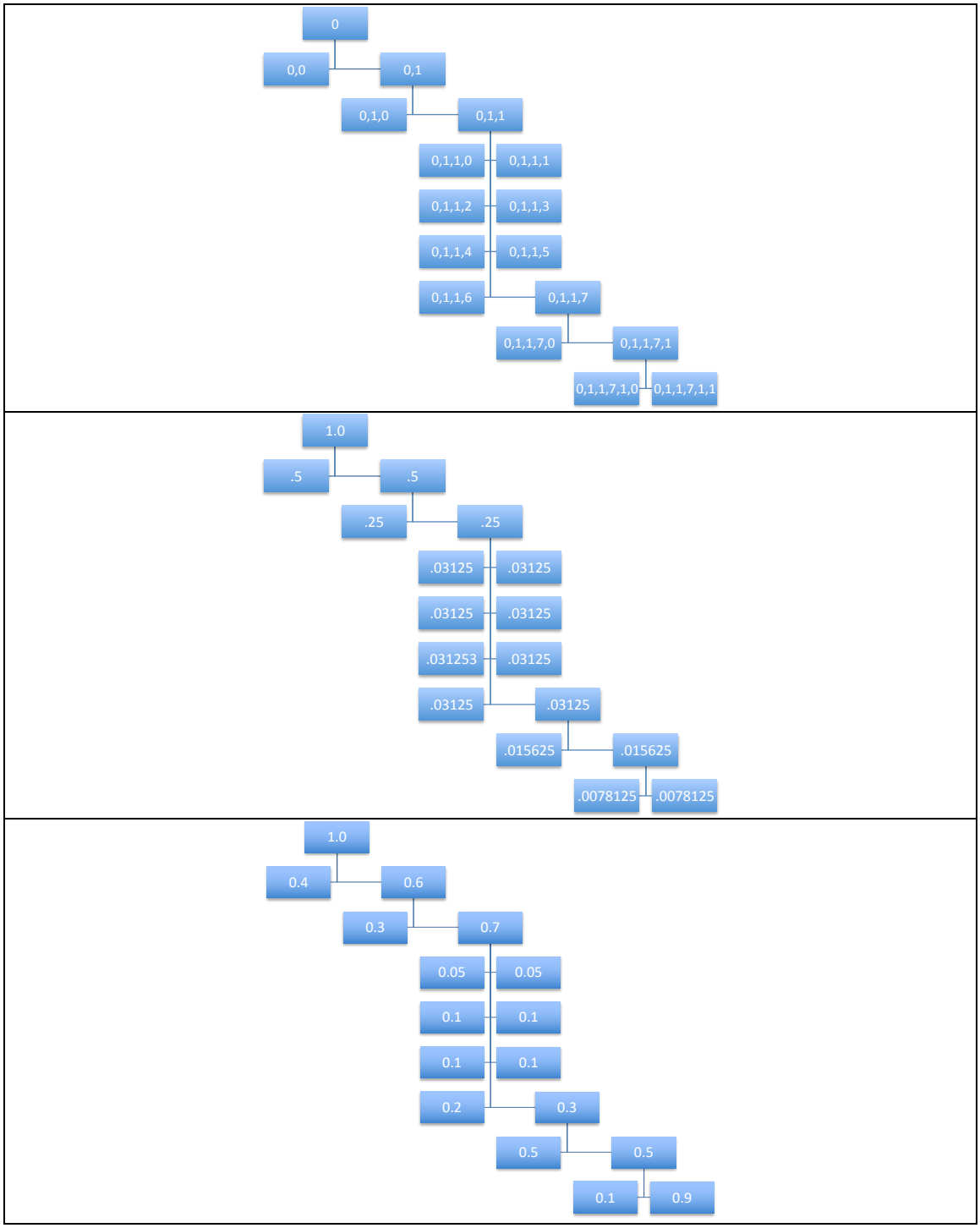


Figure 1 Label space (top), label support (middle), and label probabilities (bottom).

For the example shown in Figure 1 there are L = 5 levels where only those classes along the more probable branches of the tree are shown. Let $\omega(m)$ and $p(m)$ be the maximum likelihood class, and likelihood value at level m and $q(m)$ be the error of making a wrong decision at that level. For example, at level 1, the probability of error or making the wrong decision, i.e., that the maximum likelihood class is wrong, is

$$q(1) = 1 - p(1) = \ 0.4.$$

At level 2, the probability of making the wrong decision

$$q(2) = q(1) + [1 - q(1)][1 - p(2)] = 0.4 + [1 - 0.4][1 - 0.7] = 0.58$$

is the probability of making the wrong decision at the previous level plus the probability of making the wrong decision at this level.

Generalizing the total probability of making a wrong decision up to and including level n for $n > 0$ is

$$q(n) = \sum_{m=1}^{n} q(m-1) + [1 - q(m-1)][1 - p(m)]$$

where $p(0) = 1$ and $q(0) = 0$.

As we descend in the tree, classes become more specific and have geometrically decreasing support. If there are $L(n)$ classes at level n, their individual support is

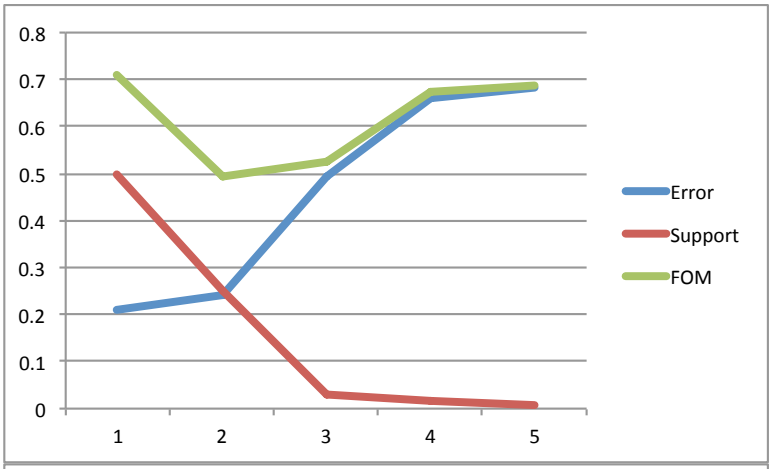
$$a(n) = \prod_{m=1}^{n} 1/L(m)$$

Let us define an objective function that combines the probability of making a wrong decision with the specificity of the decision

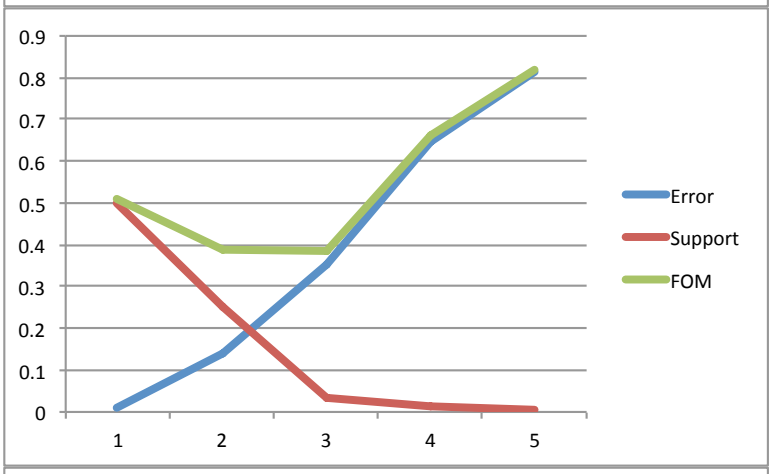
$$\lambda(n) = q(n) + a(n)$$

We define the "best" level n^* as the level that minimizes the objective function, i.e., the level with the lowest support (greatest specificity) and lowest error.

Error	Support	FOM	Max. Like.
0.21	0.5	0.71	0.79
0.2416	0.25	0.4916	0.96
0.4919	0.0312	0.5231	0.67
0.6596	0.0156	0.6752	0.67
0.6834	0.0052	0.6886	0.93



Error	Support	FOM	Max. Like.
0.01	0.5	0.51	0.99
0.1387	0.25	0.3887	0.87
0.354	0.0312	0.3853	0.75
0.6447	0.0156	0.6603	0.55
0.8117	0.0052	0.8169	0.53



Error	Support	FOM	Max. Like.
0.38	0.5	0.88	0.62
0.6838	0.25	0.9338	0.51
0.7376	0.0312	0.7688	0.83
0.7454	0.0156	0.7611	0.97
0.8791	0.0052	0.8843	0.47

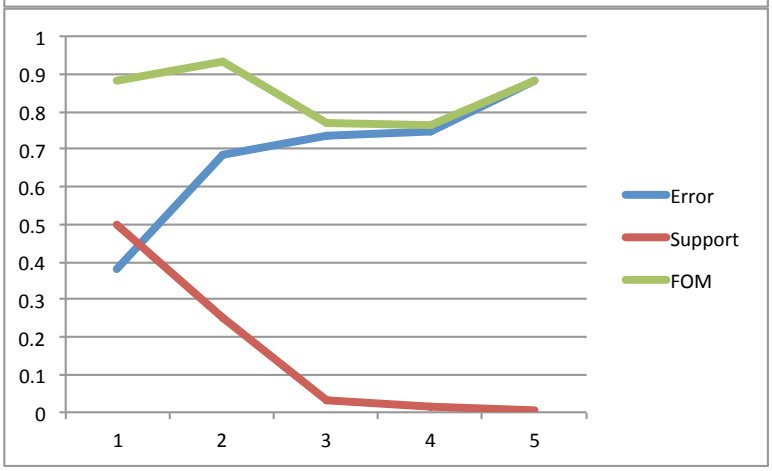


Figure 2 Simulated results. Best level (minimum FoM in graphs) highlighted in tables.

Examples

We simulated the $L = 5$ level hierarchical classifier (Figure 1) using independent random number generators that produce uniform distributions of random numbers

between zero and one. Three realizations are shown in Figure 2. The best level is highlighted in yellow in the table, which corresponds to the minimum FOM in the graph.

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.25	0.5	0.75	0.75	2
2	0.4375	0.125	0.5625	0.75	4
3	0.5781	0.015625	0.59375	0.75	8
4	0.6836	0.0009766	0.6845703	0.75	16
5	0.7627	3.052E-05	0.7627258	0.75	32

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.1	0.5	0.6	0.9	2
2	0.19	0.125	0.315	0.9	4
3	0.271	0.015625	0.286625	0.9	8
4	0.3439	0.0009766	0.3448766	0.9	16
5	0.4095	3.052E-05	0.4095405	0.9	32

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.01	0.5	0.51	0.99	2
2	0.0199	0.125	0.1449	0.99	4
3	0.0297	0.015625	0.045326	0.99	8
4	0.0394	0.0009766	0.0403806	0.99	16
5	0.049	3.052E-05	0.0490405	0.99	32

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	1E-04	0.5	0.5001	0.9999	2
2	0.0002	0.125	0.1252	0.9999	4
3	0.0003	0.015625	0.015925	0.9999	8
4	0.0004	0.0009766	0.0013765	0.9999	16
5	0.0005	3.052E-05	0.0005304	0.9999	32

Figure 3 Best levels for different simulated levels of CNN performance (max like).

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.01	0.1	0.11	0.99	10
2	0.0199	0.01	0.0299	0.99	10
3	0.0297	0.001	0.030701	0.99	10
4	0.0394	0.0001	0.039504	0.99	10
5	0.049	0.00001	0.04902	0.99	10

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.01	0.2	0.21	0.99	5
2	0.0199	0.04	0.0599	0.99	5
3	0.0297	0.008	0.037701	0.99	5
4	0.0394	0.0016	0.041004	0.99	5
5	0.049	0.00032	0.04933	0.99	5

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.01	0.3333333	0.3433333	0.99	3
2	0.0199	0.1111111	0.1310111	0.99	3
3	0.0297	0.037037	0.066738	0.99	3
4	0.0394	0.0123457	0.0517497	0.99	3
5	0.049	0.0041152	0.0531252	0.99	3

Level	Error	Support	FOM	Max. Like.	N
0	0	1		0	
1	0.01	0.5	0.51	0.99	2
2	0.0199	0.25	0.2699	0.99	2
3	0.0297	0.125	0.154701	0.99	2
4	0.0394	0.0625	0.101904	0.99	2
5	0.049	0.03125	0.08026	0.99	2

Figure 4 Best levels for different branching factors (N).

Further experiments reveal that the best class level tends to increase (classification becomes more detailed) as CNN performance increases (Figure 3). We also find as classification trees become wider, the best class level tends to decrease (become less detailed) as shown in Figure 4.

Discussion

There are two potential problems in simply selecting the level with the highest-class likelihood. If it is at a low (coarse) level, more specific class information may be ignored even if it has a relatively high likelihood. Conversely, if it is at a high (fine) level, the specificity of the class call could be misleading to the user. Although, in general, the behavior of our method is in accord with intuition, we are exploring heuristics to address cases with shallow FOMs where it may be appropriate to select a higher-level class call.

References

Chaves, Pedro (2018) "Hierarchical Classification – a useful approach for predicting thousands of possible categories," https://www.kdnuggets.com/2018/03/hierarchical-classification.html

Wu, Cinna and Tygert, Mark (2019) "A hierarchical loss and its problems when classifying non-hierarchically," https://arxiv.org/abs/1709.01062