

1 Feature score justification

One way of justifying the unbalanced correlation score is to consider the definition of entropy in information theory. Entropy measures the unlikeliness that an event will occur (the new information brought when that event actually happens). The entropy is computed as $-p_i \ln(p_i)$, where p_i is the probability of appearance of event “i”. This definition of entropy is similar to the ones humans have about noticeable events, i.e. the rarer an event is the more informative it is. We will now show that the unbalanced correlation score (for this particular problem) selects the features among those that present the lowest probability of appearance (in some sense).

In order to compute the entropy of a given feature we will need to assign to each one of them a probability measure. The features in this problem are vectors filled with ones and zeros, so there is only a finite number of possible combinations and each feature can be assigned with a probability of random appearance. A feature with a low probability of random appearance is unlikely to be randomly generated and it so may be more likely to describe some underlying input-output relation. The probability of randomness of an observation with a unbalanced score of $N = N_p - N_n$, where N_p and N_n are, respectively, the number of one entries associated to class +1 and class -1 in the feature, is:

$$P_1(T_p, T_n, N_p, N_n) = \binom{N_p + N_n}{N_p} \prod_{i=0}^{N_p+N_n-1} \frac{1}{T_p + T_n - i} \prod_{i=0}^{N_p-1} (T_p - i) \prod_{i=0}^{N_n-1} (T_n - i) \quad (1)$$

where T_p and T_n are, respectively, the total number of positive and negative labels in the training set. The combinatorial term accounts for the different number of possible combinations of appearance of N_p observations correlated with positive labels out of a feature with $N_p + N_n$ ones. The products terms accounts for the probability of one of such appearances. This probability can also be computed as:

$$P_1(T_p, T_n, N_p, N_n) = (N_p + N_n)! \binom{T_p}{N_p} \binom{T_n}{N_n} \prod_{i=0}^{N_p+N_n-1} \frac{1}{T_p + T_n - i}$$

The difference between the two is the way each observation is considered: either as a group with T_p positives and T_n negatives (the first one) or each feature individually (the second).

This probability can be used to compare features that add up to the same value in the same N , but, in order to compare the probability of appearance of features that add up to a different value, we will need to compute the probability that a certain N might occur randomly. This probability is given by:

$$P_2(T_p, T_n, N) = \frac{1}{T_p + T_n}$$

	Feature UCS	Feature Entr.
1	79650	79650
2	90405	90405
3	91838	29152
4	3391	91838
5	27150	16793
6	29152	27150
7	23405	13358
8	24797	26237
9	26237	29530
10	26264	3391
11	26952	38767
12	29530	135816
13	31643	26952
14	88024	88024
15	90587	90587
16	135816	3340

Table 1: Features selected by the unbalanced correlation score and the entropy criterion.

$$\sum_{i=0}^{\min(T_p - N, T_n - N)} P_1(T_p, T_n, \max(0, N) + i, \max(0, -N) + i)$$

Finally, one can compute the entropy for each feature as: $-P_1 P_2 \log(P_1 P_2)$, and select those with the highest values.

One can readily understand that the entropy and unbalanced score are not likely to reach the same features in any setting, because the unbalanced correlation score will not select samples with low negative N that the entropy criterion will. But in this particular problem, they do reach a similar ranking of the features, due to the unbalanced nature of the data set (40 positive examples out of 1276), the entropy criterion will tend to give high scores to those features with as many +1 samples as possible and as few -1 as possible. But for the same score N it will select those features with more positive entries. In Table 1, we show the first 16 features for both scores, where 12 of the selected features are coincident between the two methods. The 16 features for the unbalanced correlation score score either 14 (the first 3), 13

N Fs.	2	3	4	5	6	7	8	9
Tr.	0.843	0.843	0.866	0.867	0.809	0.809	0.809	0.809
Ts.	0.447	0.447	0.507	0.613	0.587	0.587	0.587	0.587

N Fs.	10	11	12	13	14	15	16
Tr.	0.809	0.809	0.821	0.868	0.868	0.868	0.868
Ts.	0.619	0.619	0.575	0.559	0.578	0.577	0.566

Table 2: Results for the the entropy criterion, features 2 to 16 using an OR function as the classifier.

(the next 3) or 12. The 4 samples that are different in the entropy criterion score to the unbalanced score 11 (feature 16793) and 10 (the other 3). These examples contain a larger number of +1s, the examples that score 10 have, respectively, 21, 20 and 19, +1 examples (and 11, 10 and 9, -1 samples).

Test success results for the entropy criterion are shown in Table 2. In the experimental results for the unbalanced correlation score we saw that the results are better for the unbalanced correlation score than for the entropy criterion, though the features they select are quite similar. We do not fully understand this.