Hill-Climbing Ensemble Feature Selection with a Larger Ensemble

David Gelbart

TR-09-001

February, 2009

Abstract

This report follows up on the investigation of hill-climbing ensemble feature selection for multi-stream speech recognition in David Gelbart's Ph.D. thesis. We compare using three streams as in the thesis to using five streams, to investigate the hypothesis that hill-climbing would give larger accuracy gains with more streams due to there being more feature selection options available. To move from three streams to five streams we split the MFCC and PLP streams into separate streams of static and dynamic features. We find that hill-climbing with five streams does result in better word recognition accuracy than three streams. This may be because our hypothesis is correct, but it also might be because the initial five stream systems that hill-climbing started from were more accurate than the initial three stream systems.

Contents

Hill-climbing ensemble feature selection with a larger ensemble

1.1 Introduction

This technical report is a follow-up to [1], which investigated the use of ensemble feature selection (EFS) for multi-stream speech recognition. In the experiments with hill-climbing EFS in [1], multi-stream ASR was always done using three classifiers (three streams). We noted in [1] that if there were more classifiers in the ensemble, the hill-climbing algorithm would have more options to choose from when distributing features. We speculated that this might result in greater accuracy gains from hill-climbing. In this report we investigate that hypothesis using five streams.

This report is meant to be read as an accompaniment to [1], rather than on its own. In particular, this report is written with the assumption that the reader is quite familiar with Chapter 6 (Hill-climbing feature selection for the Numbers task) of [1]. The experimental setup is the same apart from the change in the number of streams. We use the same hill-climbing algorithm and OGI Numbers speech recognition testbed. We control acoustic model sizes the same way, by adjusting the number of multi-layer perceptron (MLP) hidden layer units. We also do statistical significance testing the same way. (We plan to submit an INTERSPEECH 2009 conference paper based on [1] and this report. Reading the paper might be convenient as a more concise starting point than [1].)

In [1], we tried initializing hill-climbing with either randomly and non-randomly chosen initial features, and guiding hill-climbing with either ensemble accuracy or Opitz's scoring formula (Equation 5.1 in [1]). We found that non-random initialization and Opitz's formula worked the best, so in our five stream experiments we always do hill-climbing that way. As in [1] we set $\alpha = 1$ in Opitz's formula.

Our initial five feature vectors in the five-stream hill-climbing experiments were the 13 static MFCC features, the 26 dynamic (delta and delta-delta) MFCC features, the 13 static PLP features, the 26 dynamic PLP features, and the 28 MSG features. Splitting PLP features into static and dynamic streams each used by a different MLP

was previously explored in [2][3], but it worked better for us, possibly because we included delta-deltas or because we used a larger total number of streams.

As mentioned in [1], after a number of experiments were already complete, we noticed that we had used a suboptimal bunch size for MLP training. Details about this are given in Appendix A of [1]. So for consistency and comparability, in [1] and in this report we always used the same suboptimal bunch size. Similarly, we used the same duration modeling here as in [1], even though we now know a better way to do it (for more information, see the README_DURATION file we have included with our speech recognition scripts [4]).

1.2 Hill-climbing results

Table 1.1 shows basic hill-climbing results. Hill-climbing with five streams gives a statistically significant ($P < 0.01$) performance improvement over the initial five-stream features, for both clean and noisy data. Furthermore, using five streams gives both lower initial word error rates (WERs) and lower final WERs than in any of the three-stream rows in Table 1.1.

Comparing the five-stream system to the three-stream systems with non-random initial choice of features, the difference in initial WERs is statistically significant for clean data ($P < 0.003$) but not for noisy data. Comparing to the three-stream systems with random initial choice of features, the difference in initial WERs is statistically significant for both clean ($P < 0.02$) and noisy ($P < 10^{-7}$) data.

For the clean data, the difference in final WERs is not statistically significant for row (d) of Table 1.1 but it is statistically significant for rows (a) ($P < 0.008$), (b) ($P < 0.004$), and (c) ($P = 0.01$). For the noisy data, the difference in final WERs is statistically significant in all four cases ($P < 0.003$).

In the clean case there was no statistically significant difference between the initial five-stream system and the final systems chosen by three-stream hill-climbing. In the noisy case the final three stream systems were all better than the initial five-stream system ($P < 0.0003$).

Row (c) in Table 1.1 is the closest three-stream counterpart to the five-stream hill-climbing experiment, since both use $\alpha = 1$ with non-random initial choice of features. In row (c) the relative WER improvement from hill-climbing, relative to the initial WER, is 8.2% for the clean data and 5.7% for the noisy data. In the five-stream case it is 6.7% for the clean data and 9.0% for the noisy data. For the clean data, the initial five stream system had a substantially lower WER than the initial three stream system, which may be the reason why the relative improvement from hill-climbing was smaller.

Table 1.1 shows that many more changes were made to feature vectors during hill-climbing with five streams than with three. Accordingly, the time taken to run hill-climbing was also longer with five streams, due to more passes through the feature

Experiment	Clean train and test			Noisy train and test		
	Changes	Initial WER	Final WER	Changes	Initial WER	Final WER
(a) Hill-climbing (HC), 3 streams	4	4.9	4.6	5	15.7	14.8
(b) HC, 3 streams, RSM initialization	2	4.8	4.6	17	17.2	14.8
(c) HC, 3 streams, using $\alpha = 1$	14	4.9	4.5	20	15.7	14.8
(d) HC, 3 streams, using $\alpha = 1$, RSM initialization	17	4.8	4.4	15	17.2	14.9
(e) HC, 5 streams, using $\alpha = 1$	45	4.5	4.2	61	15.6	14.2

Table 1.1. This table shows hill-climbing results on the Numbers corpus. The "Changes" columns give the number of features changed (added to or deleted from a feature vector) during hill-climbing. The "Initial WER" columns give the initial ensemble (i.e., using all streams) word error rate on the evaluation set before the hill-climbing algorithm has made any changes to the feature vectors. The "Final WER" columns give the ensemble word error rate on the evaluation set once hill climbing has finished. If a value is given for α it means that the score used to guide hill-climbing was the formula defined in Equation 5.1 of [1], calculated on the development set. Otherwise, the score used to guide hill-climbing was the ensemble WER on the development set. "RSM" means that initial feature vectors prior to the start of hill-climbing were chosen randomly. Otherwise, the initial feature vectors were respectively MFCC, PLP and MSG for three streams or static MFCC features, dynamic MFCC features, static PLP features, dynamic PLP features and MSG for five streams.

set and more repeated trainings (see section 6.4 of [1]). Five-stream hill-climbing took about 52 days in the clean case and 91 days in the noisy case. In the noisy case, we sped up hill-climbing for the last two streams by using an 8-core machine for decodings instead of a 4-core machine as in [1]. As discussed in [1], it might have been possible to speed things up by doing decoder parameter tuning less often.

1.3 Additional baseline results

Table 1.2 shows additional baseline results. In the clean case, row (d) (using all features in the feature pool concatenated together into a single feature vector) and row (g) (five streams) are tied for the lowest WER. There is a statistically significant difference ($P < 0.02$) between those two systems and the system with the second lowest WER in row (f).

In the noisy case, row (d) has the lowest WER, and row (g) has the second lowest

WER. There is a statistically significant difference between the two ($P < 10^{-4}$). There is no statistically significant difference between the system in row (g) and the system with the third lowest WER in row (e).

Experiment	Clean train and test	Noisy train and test
(a) MFCC	6.5	21.4
(b) PLP	5.0	17.5
(c) MSG	7.3	16.3
(d) All features concatenated	4.5	14.7
(e) MFCC, PLP, MSG (three MLPs)	4.9	15.7
(f) RSM (three MLPs)	4.8	17.2
(g) Static MFCC, dynamic MFCC, static PLP, dynamic PLP, MSG (five MLPs)	4.5	15.6

Table 1.2. This table shows baseline WERs on the Numbers corpus. In row (d), all features in the feature pool are concatenated into a single feature vector. In row (e), there are three MLPs respectively using 39 MFCC features (1200 MLP hidden units), 39 PLP features (1200 HUs), and 28 MSG features (1590 HUs). Row (e) corresponds to the "Initial WER" column in rows (a) and (c) of Table 1.1. In row (f), there are three MLPs, using randomly chosen feature vectors with the same number of features and hidden units as in row (e). Row (f) corresponds to the "Initial WER" column in rows (b) and (d) of Table 1.1. In row (g) there are five MLPs respectively using 13 static MFCC features (1716 HUs), 26 dynamic MFCC features (1014 HUs), 13 static PLP features (1716 HUs), 26 dynamic PLP features (1014 HUs), and 28 MSG features (954 HUs). Row (g) corresponds to the "Initial WER" column in row (e) of Table 1.1.

1.4 Hill-climbing results compared to the additional baseline results

In the clean case, the final system chosen by five-stream hill-climbing has lower WER than any of the baselines in Table 1.2. The difference between that and the two best baselines (the five-stream initial system and the all-features-concatenated baseline) is statistically significant ($P < 0.01$ for the five-stream baseline and $P < 0.04$ for the all-features-concatenated baseline).

In the noisy case, the difference between the final system chosen by five-stream hill-climbing and the best baseline in Table 1.2, the all-features-concatenated baseline, is not statistically significant ($P = 0.052$).

1.5 The feature vectors chosen by hill-climbing

To view diagrams showing the initial and final feature vectors in each hill-climbing experiment, visit `http://www.icsi.berkeley.edu/speech/papers/gelbart-ms/numbers-hillclimb`. You can also download the lists of initial and final features in plain text from that location.

1.6 Hill-climbing progress over time

Figures 1.1 and 1.2 plot the progress of the hill-climbing algorithm over time for the five-stream experiments. The points labeled as "Initial" in the figures correspond to the initial feature vectors before any changes were made. The other points correspond to changes made to feature vectors and are labeled according to what stream was changed.

Since we used Equation 5.1 (Opitz's formula) from [1] to guide the hill-climbing process, the figures show Equation 5.1 for the development set, development set ensemble word recognition accuracy, and evaluation set ensemble word recognition accuracy. We plot ensemble word recognition accuracy instead of ensemble WER in this case because the former is easier to plot on the same graph as Equation 5.1. Word recognition accuracy is simply 100% minus the WER.

The values of Equation 5.1 labeled "Initial" in the figures were calculated for stream 1 using the initial feature vectors for each stream. Since the value of Equation 5.1 depends on the current stream (since the equation is based on that stream's accuracy and its contribution to ensemble diversity), it's possible for it to get worse when the hill-climbing algorithm moves from one stream to the next.

In the clean case (Figure 1.1), the ways that evaluation set ensemble accuracy and development set ensemble accuracy changed over time were fairly similar. Thus, although the final value of the evaluation set ensemble accuracy is lower than the peak value, the figure does not imply that this was caused by overfitting [5][6][7][8] to the development set.

In the noisy case (Figure 1.1), on the other hand, evaluation set ensemble accuracy reached a peak and then decreased between changes 40 and 60, but the development set ensemble accuracy did not do the same. This suggests overfitting to the development set may have occurred.

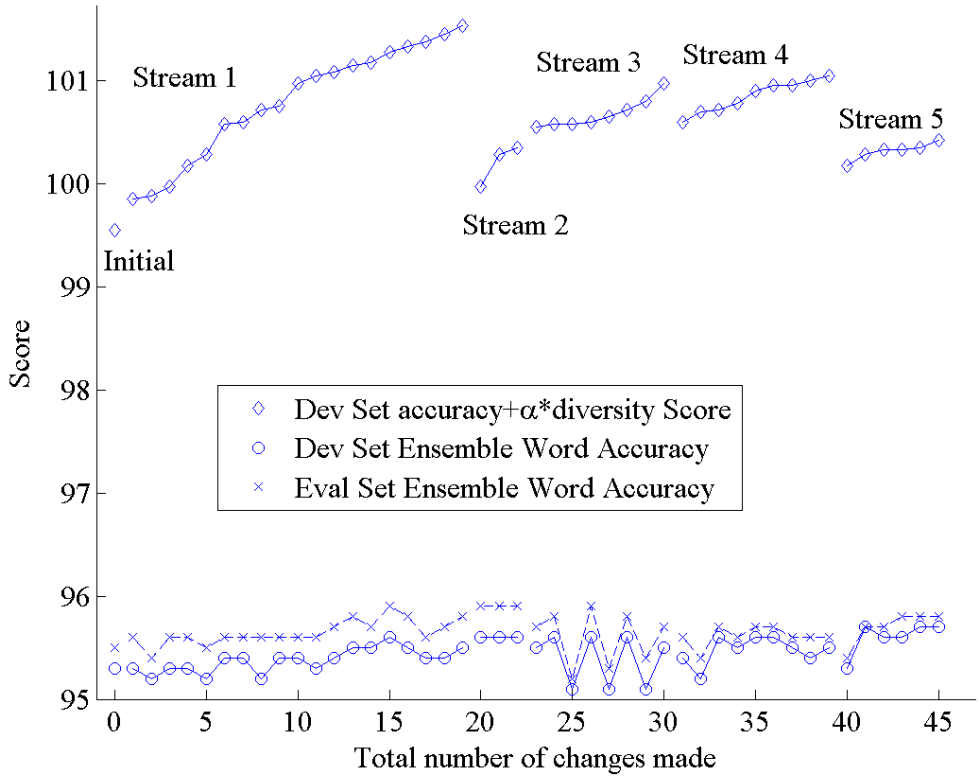

Figure 1.1. Progress of the hill-climbing algorithm over time for the clean Numbers data.

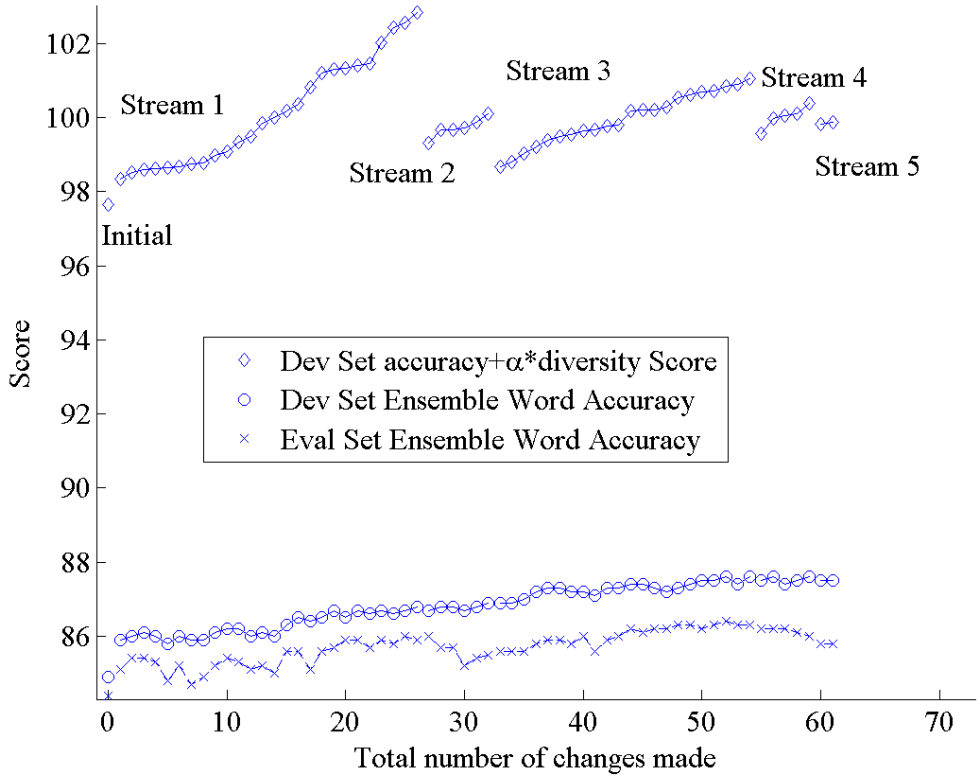

Figure 1.2. Progress of the hill-climbing algorithm over time for the noisy Numbers data.

1.7 Hill-climbing performance for seen vs. unseen noises

In our noisy version of the Numbers corpus there are four noise types that are used in the training set, the development set, and the evaluation set. There are also two noises that are only in the training set, two that are only in the development set, and two that are only in the evaluation set.

This raises the question of whether hill-climbing improved evaluation set performance for all six noise types in the evaluation set, or only for the four shared noise types. Table 1.3 compares evaluation set WERs for the four shared noises and the two evaluation-only noises.

For the shared noises, the difference between initial and final WERs for five-stream hill-climbing is statistically significant ($P < 10^{-17}$). However, the difference between the initial and final WERs for evaluation-only noises for five-stream hill-climbing is not statistically significant. This differs from the three-stream case in which hill-climbing improved on initial performance even for the evaluation-only noises [1]. This may be because the initial five-stream WER for evaluation-only noises is 22.4% which is considerably lower than the initial three-stream WERs for evaluation-only noises.

Experiment	Shared noises	Evaluation-only noises
(a) Initial 3 streams	14.1	25.5
(b) Initial 3 streams, chosen by RSM	15.0	28.8
(c) Hill-climbing (HC), 3 streams	13.3	23.4
(d) HC, 3 streams, with RSM initial streams	13.3	23.6
(e) HC, 3 streams, using $\alpha = 1$	13.1	24.2
(f) HC, 3 streams, using $\alpha = 1$, with RSM initial streams	13.1	24.5
(g) Initial 5 streams	15.0	22.4
(h) HC, 5 streams, using $\alpha = 1$	12.7	22.7

Table 1.3. Word error rates on the evaluation set for the noisy Numbers corpus, divided into shared noises and evaluation-only noises. The WERs for shared noises were calculated over 1970 utterances (9821 words). The WERs for evaluation-only noises were calculated over 986 utterances (4787 words). There were also 592 evaluation set utterances without any noise added.

1.8 Testing the features chosen by hill climbing in heavily mismatched conditions

We now investigate how the features chosen by hill-climbing perform if we use them in a condition (clean or noisy) that differs from the situation during hill-climbing.

First, we present results for which we used the clean Numbers corpus for hill-climbing, MLP training and decoder parameter tuning, but tested on the evaluation set from the noisy Numbers corpus. Second, we present results for which we used the noisy Numbers corpus for hill-climbing, but then calculated final evaluation set results by doing MLP training and decoder parameter tuning using the clean corpus, and then testing on the evaluation set from the noisy corpus.

1.8.1 Using the noisy evaluation set with the clean training and development sets

Table 1.4 shows the results of using the clean Numbers corpus for hill-climbing, MLP training and decoder parameter tuning but testing on the noisy Numbers evaluation set. The best result was obtained by the initial five-stream system. There is not a statistically significant difference between that and the second-best result, which was obtained by the final hill-climbing system in row (c). Both the final system in row (c) and the initial five-stream system in row (e) performed better than the final five-stream system chosen by hill-climbing, and the difference is statistically significant ($P < 0.008$). However, the final five-stream system chosen by hill-climbing performed better than the systems in the remaining rows (a, b and d) of the table, and this is statistically significant ($P < 10^{-26}$).

Table 1.5 shows the corresponding baseline results. The best result, by far, was obtained using five streams.

Experiment	Initial WER	Final WER
(a) Hill-climbing (HC), 3 streams	30.5	29.0
(b) HC, 3 streams, RSM initialization	30.5	28.2
(c) HC, 3 streams, using $\alpha = 1$	30.5	23.4
(d) HC, 3 streams, using $\alpha = 1$, RSM init.	30.5	27.9
(e) HC, 5 streams, using $\alpha = 1$	22.9	24.1

Table 1.4. Hill-climbing WER results on the evaluation set in a highly mismatched condition: the clean training and development sets were used with the noisy evaluation set. "Initial WER" refers to the WER for the initial features that the hill-climbing process started with. "Final WER" refers to the WER of the final features chosen by hill-climbing.

Experiment	WER
(a) MFCC	33.2
(b) PLP	38.0
(c) MSG	30.4
(d) All features concatenated	32.1
(e) MFCC, PLP, MSG (three MLPs)	30.5
(f) RSM (three MLPs)	30.5
(g) Static MFCC, dynamic MFCC, static PLP, dynamic PLP, MSG (five MLPs)	22.9

Table 1.5. Baseline WERs on the evaluation set in a highly mismatched condition: the clean training set was used with the noisy evaluation set. The feature vectors and MLP hidden layer sizes are the same as in Table 1.2.

1.8.2 Using features from hill-climbing on noisy data with the clean training set and the noisy evaluation set

Table 1.6 shows the results of using features selected by hill-climbing for the noisy data, but then calculating evaluation set WER by performing MLP training and decoder tuning on the clean training set and then testing on the noisy evaluation set.

The best WER is from the initial five-stream system and the second best WER is from the final five-stream system. The difference between the two is statistically significant ($P < 10^{-28}$). The difference between the second best WER and the third best WER, which is from the final three-stream system in row (c), is also statistically significant ($P < 10^{-7}$).

The final WER in each row of Table 1.6 is worse than the final WER in the corresponding row of Table 1.4 and this is statistically significant in each case ($P < 10^{-6}$ in rows (a), (b), (c) and (e); $P = 0.03$ in row (d)).

Experiment	Initial WER	Final WER
(a) Hill-climbing (HC), 3 streams	30.5	30.5
(b) HC, 3 streams, RSM initialization	30.5	31.9
(c) HC, 3 streams, using $\alpha = 1$	30.5	28.2
(d) HC, 3 streams, using $\alpha = 1$, RSM init.	30.5	28.5
(e) HC, 5 streams, using $\alpha = 1$	22.9	26.7

Table 1.6. Hill-climbing WER results on the evaluation set in a highly mismatched condition: hill-climbing was performed using the noisy data, but evaluation set WER was measured by performing MLP training and decoder tuning on the clean training set and then testing on the noisy evaluation set.

1.9 Summary and Conclusions

Overall, five streams (i.e., five MLPs in an ensemble) performed better than three streams. This was true both when comparing the three- and five-stream initial systems and when comparing the systems chosen by hill-climbing (Table 1.1). Feature concatenation performed better overall than the initial five-stream system (Table 1.2), but the five-stream system chosen by hill-climbing performed better overall than feature concatenation (Tables 1.1 and 1.2).

Furthermore, for both clean and noisy data the five-stream system chosen by hill-climbing delivered a statistically significant improvement in WER over the five-stream initial system (Table 1.1). In Figures 1.1 and 1.2 we plotted the progress of the hill-climbing algorithm over time, from which we learned that overfitting to the development set may have lowered final five-stream performance for the noisy data. Thus it may be possible to attain a larger improvement over the five-stream initial system by addressing the issue of overfitting.

However, when we divided the noisy Numbers results into shared vs. evaluation-only noises (Table 1.3), we found that for the evaluation-only noises, which were not seen during hill-climbing, there was no statistically significant difference between the initial and final five-stream systems. On the other hand, the three-stream final systems were better than the three-stream initial systems for the evaluation-only noises, possibly because the three-stream initial systems were easier to improve on due to higher WERs.

Later, when we used the clean Numbers corpus for hill-climbing and MLP training but measured final performance on the noisy Numbers evaluation set, we found that the initial five-stream system and one of the final three-stream systems outperfomed the other systems including the final five-stream system (Tables 1.4 and 1.5). However, the final five-stream system outperformed the remaining systems.

When we used features selected by hill-climbing for the noisy data, but then measured final performance by performing MLP training and decoder tuning on the clean training set and then testing on the noisy evaluation set (Table 1.6), we found that the final five-stream system was better than the three-stream systems, but the initial five-stream system was better than the final five-stream system.

It is hard to say whether the stronger overall performance of the systems chosen by hill-climbing using five streams rather than than three streams is because five streams provides more feature selection options (as we speculated in the Introduction) or simply because the initial five-stream system hill-climbing started from generally performed better than the initial three-stream systems. Of course, it could be that both played a role, or that there are other important factors that we are overlooking.

Bibliography

[1] D. Gelbart, *Ensemble Feature Selection for Multi-Stream Automatic Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 2008, online at `http://www.icsi.berkeley.edu`.

[2] D. Ellis, "Stream combination before and/or after the acoustic model," Tech. Rep. 00-007, International Computer Science Institute, 2000, available at `http://www.icsi.berkeley.edu`.

[3] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," in *ICSLP*, Beijing, China, 2000.

[4] "Noisy Numbers and Numbers ASR scripts," `http://www.icsi.berkeley.edu/Speech/papers/gelbart-ms/numbers`.

[5] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, vol. 3, 2003.

[6] J. Loughrey and P. Cunningham, "Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search," Tech. Rep. 2005-37, Trinity College Dublin, 2005, available at `http://www.cs.tcd.ie`.

[7] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets," Tech. Rep. 2005-17, Trinity College Dublin, 2005, available at `http://www.cs.tcd.ie`.

[8] P. Cunningham, "Overfitting and diversity in classification ensembles based on feature selection," Tech. Rep. 2000-07, Trinity College Dublin, 2000, available at `http://www.cs.tcd.ie`.