In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 103–106, Lisbon, Portugal, 2000.

Genetic Algorithms for Feature Relevance Assignment in Memory-Based Language Processing

Anne Kool and **Walter Daelemans** and **Jakub Zavrel***
CNTS – Language Technology Group
University of Antwerp, UIA, Universiteitsplein 1, 2610 Antwerpen, Belgium
{kool, daelem, zavrel}@uia.ua.ac.be

Abstract

We investigate the usefulness of evolutionary algorithms in three incarnations of the problem of feature relevance assignment in memory-based language processing (MBLP): feature weighting, feature ordering and feature selection. We use a simple genetic algorithm (GA) for this problem on two typical tasks in natural language processing: morphological synthesis and unknown word tagging. We find that GA feature selection always significantly outperforms the MBLP variant without selection and that feature ordering and weighting with GA significantly outperforms a situation where no weighting is used. However, GA selection does not significantly do better than simple iterative feature selection methods, and GA weighting and ordering reach only similar performance as current information-theoretic feature weighting methods.

1 Memory-Based Language Processing

Memory-Based Language Processing (Daelemans, van den Bosch, and Zavrel, 1999) is based on the idea that language acquisition should be seen as the incremental storage of exemplars of specific tasks, and language processing as analogical reasoning on the basis of these stored exemplars. These exemplars take the form of a vector of, typically, nominal features, describing a linguistic problem and its context, and an associated class symbol representing the solution to the problem. A new instance is categorized on the basis of its similarity with a memory instance and its associated class.

The basic algorithm we use to calculate the distance between two items is a variant of IB1 (Aha, Kibler, and Albert, 1991). IB1 does not solve the problem of modeling the difference in relevance between the various sources of information. In an MBLP approach, this can be overcome by means of *feature weighting.* The IB1-IG algorithm uses information gain to weight the cost of a feature value mismatch during comparison. IGTREE is a variant in which an oblivious decision tree is created with features as tests, and in which tests are ordered according to information gain of the associated features. In this case, the accuracy of the trained system is very much dependent on a good *feature ordering.* For all variants of MBLP discussed here, *feature selection* can also improve both accuracy and efficiency by discarding some features altogether because of their irrelevance or even counter-productivity in learning to solve the task. In our experiments we will use a relevance assignment method that radically differs from information-theoretic measures: genetic algorithms.

2 Genetic Algorithms for Assigning Relevance

In the experiments, we linked our memory-based learner TIMBL[1] to PGAPACK[2]. During the weighting experiments a gene corresponds to a specific real-valued feature-weight (we will indicate this by including GA in the algorithm name, i.e. IB1-GA and GATREE, cf. IB1-IG and IGTREE).

* Research funded by CELE, S.AI.L Trust V.Z.W., Ieper, Belgium.

[1] TIMBL is available from http://ilk.kub.nl/ and the algorithms are described in more detail in (Daelemans et al., 1999).

[2] A software environment for evolutionary computation developed by D. Levine, Argonne National Laboratory, available from ftp://ftp.mcs.anl.gov/pub/pgapack/

In the case of selection the string is composed of binary values, indicating presence or absence of a feature (we will call this GASEL). The fitness of the strings is determined by running the memory-based learner with each string on a validation set, and returning the resulting accuracy as a fitness value for that string. Hence, both weighting and selection with the GA is an instance of a *wrapper* approach as opposed to a *filter* approach such as information gain (Kohavi and John, 1995).

For comparison, we include two popular classical wrapper methods: backward elimination selection (BASEL) and forward selection (FOSEL). Forward selection starts from an empty set of features and backward selection begins with a full set of features. At each further addition (or deletion, for BASEL) the feature with the highest accuracy increase (resp. lowest accuracy decrease) is selected, until improvement stalls (resp. performance drops).

During the morphology experiment the population size was 50, but for prediction of unknown words it was set to 16 because the larger dataset was computationally more demanding. The populations were evolved for a maximum of 200 generations or stopped when no change had occurred for over 50 generations. Parameter settings for the genetic algorithm were kept constant: a two-point crossover probability of 0.85, a mutation rate of 0.006, an elitist replacement strategy, and tournament selection.

2.1 Data

The first task[3] we consider is prediction of what diminutive suffix a Dutch noun should take on the basis of its form. There are five different possible suffix forms (the classes). There are 12 features which contain information (stress and segmental information) about the structure of the last three syllables of a noun. The data set contains 3949 such instances.

The second data set[4] is larger and contains 65275 instances, the task we consider here is part-of-speech (morpho-syntactic category) tagging of unknown words. The features used here are the coded POS-tags of two words before and two words after the focus word to be tagged, the last three letters of the focus word, and information on hyphenation and capitalisation. There are 111 possible classes (part of speech tags) to predict.

2.2 Method

We have used 10-fold-cross-validation in all experiments. Because the wrapper methods get their evaluation feedback directly from accuracy measurements on the data, we further split the trainfile for each fold into 2/3 sub-trainset and a 1/3 validation set. The settings obtained by this are then tested on the test set of that fold.

2.3 Results

In Table 1 we show the results of our experiments (average accuracy and standard deviation over ten folds). We can see that applying any feature selection scheme when no weights are used (IB1) significantly improves classification performance ($p<0.01$)[5]. Selection also improves accuracy when using the IB1-IG or IGTREE algorithm. These differences are significant on the morphology dataset ($p<0.05$), but for the unknown words dataset only the difference between (IB1) and (IB1+GASEL) is significant ($p<0.01$). In both cases, however, the results in Table 1 do not reveal significant differences between evolutionary, backward or forward selection.

With respect to feature weighting by means of a GA the results are much less clear: for the morphology data, the GA-weights significantly improve upon IB1, refered to as IB1-GA in the table, ($p<0.01$) but not IGTREE (GATREE in the table). For the other dataset GA-weights do not even improve upon IB1. But in general, those weights found by the genetic algorithm lead to comparable classification accuracy as with gain ratio based weighting. The same applies to the combination of GA-weights with further selection of irrelevant features (GATREE+GASEL).

2.4 The Effect of GA Parameters

We also wanted to test whether the GA would benefit from optimisation in the crossover and mutation probabilities. To this end, we used the morphology dataset, which was split into an 80% trainfile, a 10% validationfile and a held-out 10% testfile. The mutation rate was var-

[3]Data from the CELEX lexical data base, available on CD-ROM from the LDC, `http://ldc.upenn.edu`.

[4]This dataset is based on the TOSCA tagged LOB corpus of English.

[5]All significance tests in this paper are one-tailed paired t-tests.

Classifier	Morphology	Unknown Words
IB1	87.2 (± 1.6)	81.7 (± 0.5)
IB1+GASEL	96.5 (± 1.0)	82.8 (± 0.6)
IB1+FOSEL	96.6 (± 1.1)	82.9 (± 0.2)
IB1+BASEL	96.6 (± 1.1)	82.9 (± 0.2)
IB1-IG	96.2 (± 0.8)	82.8 (± 0.3)
IB1-IG+GASEL	**97.3** (± 0.9)	**83.0** (± 0.3)
IB1-IG+FOSEL	97.1 (± 0.9)	82.8 (± 0.3)
IB1-IG+BASEL	**97.3** (± 1.0)	82.9 (± 0.3)
IGTREE	96.2 (±0.8)	**81.4** (± 0.4)
IGTREE+GASEL	**97.1** (± 0.9)	**81.4** (± 0.4)
IGTREE+FOSEL	97.0 (± 0.9)	81.3 (± 0.4)
IGTREE+BASEL	97.0 (± 1.1)	81.3 (± 0.4)
IB1-GA	95.6 (± 1.0)	81.6 (± 0.8)
IB1-GA+GASEL	97.0 (± 1.1)	82.0 (± 1.2)
GATREE	96.0 (± 1.0)	80.4 (± 1.2)
GATREE+GASEL	**97.1** (± 1.0)	81.0 (± 0.6)

Table 1: Accuracy (± standard deviation) results of the experiments. Boldface marks the best results for each basic algorithm per data set.

ied stepwise adding a value of 0.001 at each experiment, starting at a 0.004 value up to 0.01. The different values for crossover ranged from 0.65 to 0.95, in steps of 0.O5. The effect of changing crossover and mutation probabilities was tested for IB1-IG+GA-selection, for IB1 with GA weighting, for IGTREE+GA-selection, and for IGTREE with GA-weight settings.
These experiments show considerable fluctuation in accuracy within the tested range, but different parameter settings could also yield same results although they were far apart in value. Some settings achieved a particularly high accuracy in this training regime (e.g. crossover: 0.75, mutation: 0.009). However, when we used these in the ten-fold cv setup of our main experiments, this gave a mean score of 97.4 (± 0.9) for IB1-IG with GA-selection and a mean score of 97.1 (± 1.1) for IGTREE with GA-selection. These accuracies are similar to those achieved with our default parameter settings.

2.5 Discussion

Feature selection on the morphology task shows a significant increase in performance accuracy, whereas on the unknown words task the differences are less outspoken. To get some insight into this phenomenon, we looked at the average probabilities of the features that were left out by the evolutionary algorithm and their average weights.

On the morphology task this reveals that nucleus and coda of the last syllable are highly relevant, they are always included. The onset of all three syllables is always left out. Further, in all partitions the nucleus and coda of the second syllable are left out. [6] For part-of-speech tagging of unknown words all features appear to be more or less equally relevant. Over the ten partitions, either no omission is suggested at all, or the features that carry the pos-tag of n-2 word before and the n+2 word after the focus word are deleted. This is comparable to reducing the context window of this classification task to one word before and one after the focus. The fact that all features seem to contribute to the classification when doing POS-tagging (making selection irrelevant) could also explain why the IGTREE algorithm seems to benefit less from the feature orders suggested and why the non-weighted approach IB1 already has a high score on the tagging task. The IGTREE algorithm is more suited for problems where the features can be ordered in a straightforward way because they have significantly different relevance.

3 Conclusions and Related Research

The issue of feature-relevance assignment is well-documented in the machine learning literature. Excellent comparative surveys are (Wettschereck, Aha, and Mohri, 1997) and (Wettschereck and Aha, 1995) or (Blum and Langley, 1997). Feature subset selection by means of evolutionary algorithms was investigated by Skalak (1994), Vafaie and de Jong (1992), and Yang and Honavar (1997). Other work deals with evolutionary approaches for continuous feature weight assignment such as Wilson and Martinez (1996), or Punch and Goodman (1993).

The conclusions from these papers are in agreement with our findings on the natural language data, suggesting that feature selection and weighting with GA's significantly outperform non-weighted approaches. Feature selection generally improves accuracy with a reduc-

[6] This fits in with current theory about this morphological process (e.g. Trommelen (1983), Daelemans et al. (1997)).

tion in the number of features used. However, we have found no results (on these particular data) that indicate an advantage of evolutionary feature selection approach over the more classical iterative methods. Our experiments further show that there is no evidence that GA weighting is *in general* competitive with simple filter methods such as gain ratio. Possibly, a parameter setting for the GA could be found that gives better results, but searching for such an optimal parameter setting is at present computationally unfeasible for typical natural language processing problems.

References

Aha, D., D. Kibler, and M. Albert. 1991. Instance-based learning algorithms. In *Machine Learning Vol. 6, pp 37-66.*

Blum, A. and P. Langley. 1997. Selection of relevant features and examples in machine learning. In *Machine Learning: Artificial Intelligence,97, pp 245-271.*

Daelemans, W., P. Berck, and S. Gillis. 1997. Data mining as a method for linguistic analysis: Dutch diminutives. In *Folia Linguistica , XXXI/1-2, pp 57-75.*

Daelemans, W., A. van den Bosch, and J. Zavrel. 1999. Forgetting exceptions is harmful in language learning. In *Machine Learning, special issue on natural language learning, 34 , pp 11-43.*

Daelemans, W., J. Zavrel, K. van der Sloot, and A. van den Bosch. 1999. Timbl: Tilburg memory based learner, version 2.0, reference guide. Ilk technical report 99-01, ILK.

John, G.H., R. Kohavi, and K. Pfleger. 1994. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference, pp 121-129.*

Kohavi, R. and G.H. John. 1995. Wrappers for feature subset selection. In *Artificial Intelligence Journal, Special Issue on Relevance Vol.97, pp 273-324.*

Punch, W. F., E.D. Goodman, Lai Chia-Shun Min Pei, P. Hovland, and R. Enbody. 1993. Further research on feature selection and classification using genetic algorithms. In *Proceedings of the Fifth International Conference on Genetic Algorithms, pp 557.*

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann.

Skalak, D. 1993. Using a genetic algorithm to learn prototypes for case retrieval and classification. In *Case-Based Reasoning: Papers from the 1993 Workshop, Tech. Report WS-93-01, pp 211-215. AAAI Press.*

Skalak, D. B. 1994. Prototype and feature selection by sampling and random mutation hill climbing algorithms. In *Proceedings of the eleventh International Conference on Machine Learning, pp 293-301.*

Trommelen, M.T.G. 1983. *The Syllable in Dutch, with special Reference to Diminutive Formation.* Foris: Dordrecht.

Vafaie, H. and K. de Jong. 1992. Genetic algorithms as a tool for feature selection in machine learning. In *Machine Learning, Proceeding of the 4th International Conference on Tools with Artificial Intelligence, pp 200-204.*

Wettschereck, D. and D. Aha. 1995. Weighting features. In *Proceedings of the First International Conference on Case-Based Reasoning, ICCBR-95, pp 347-358.*

Wettschereck, D., D. Aha, and T. Mohri. 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. In *Artificial Intelligence Review Vol.11, pp 273-314.*

Wilson, D. and T. Martinez. 1996. Instance-based learning with genetically derived attribute weights. In *Proceedings of the International Conference on Artificial Intelligence, Expert Systems, and Neural Networks, pp 11-14.*

Yang, J. and V. Honavar. 1997. Feature subset selection using a genetic algorithm. In *Genetic Programming 1997: Proceedings of the Second Annual Conference, pp 380.*