A Review: Decision Trees in Speech Recognition

Teemu Hirsimäki
teemu.hirsimaki@hut.fi

May 27, 2003

Abstract

Decision trees have been used widely for clustering in many research areas during the past decades. This paper reviews literature on the use of decision trees in the field of speech recognition: how they have been used to improve acoustic modelling, and what kind of language models have been built using them.

1 Introduction

During the past decades, the most popular acoustic model used in speech recognition has been the hidden Markov model. It consist of states with emission probability densities and transitions between states (see Fig. 1). The basic idea is to create a separate N-state model for each phoneme of the language, and use transcribed speech data to estimate the parameters of the model.

However, it was found that the basic model could be improved by taking the contexts of the phonemes into account. Because vocal articulators can not change shape very fast, the articulation of a phoneme is affected by the neighboring phonemes. This *co-articulation* effect is quite considerable and degrades the performance of speech recognition systems not taking the effect into account.

In order to improve the accuracy of the models, most modern speech recognition systems use so called *context-dependent* models. Instead of training a single HMM for each phoneme, several models are trained according to the context of the phoneme. In the HTK notation [Odell, 1995], the label `h-i+s` means phoneme `i` having the context `h` on the left side and `s` on the right side. For instance, if the training speech data contains many occurrences of `i` in the contexts of `h-i+s` and `a-i+m`, separate models are trained for these two cases. This kind of models using one preceding and following phoneme context are often called as *triphones*, while the models without context are called as *monophones*.

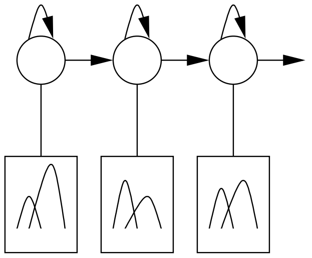


Figure 1: An example of a 3-state hidden Markov phoneme model.

Even if the context-dependent models improve the modelling accuracy, the number of the models increase heavily, and the obvious problem is that there is much less training data for each model. If the system models 40 different phonemes, the number of triphones is as high as $40^3 = 64000$, and most of the triphones are so rare, that there might be missing from the training data completely.

Decision tree is one approach to solve the problem, and there has been active research on the 90's how to improve acoustic models with them. The basic idea is to use decision tree to find similar triphones, and share the parameters between them.

In Section 2 the basic ideas behind the decision trees are explained, and Section 3 reviews literature using decision trees in acoustic modelling. In Section 4 we take a brief look at how decision trees have been used to improve language modelling, which is another important part of speech recognition systems. Finally, Section 5 summarises the work.

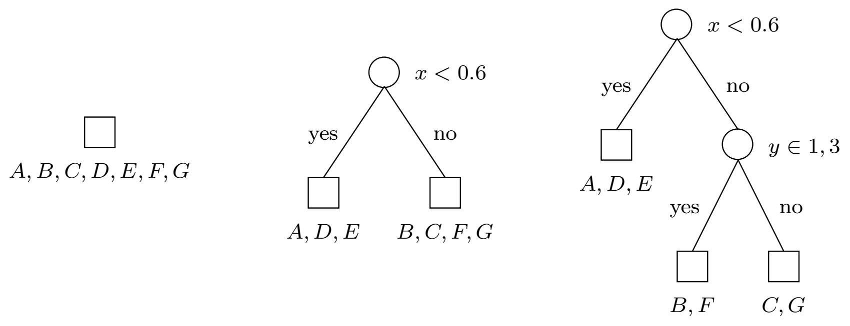
2 Decision trees

Decision tree [Breiman et al., 1984] is a popular clustering method which has been used in many research areas. It uses a top-down approach to split the samples, which are initially in one cluster, to smaller clusters by asking questions about the samples.

Figure 2 illustrates the idea. Initially, all training samples (A, B, C, D, E, F, G having properties x and y) are placed in the root node (Fig. 2b). Then a question about the properties of the samples is made, and according to the answers, new child nodes are created and the samples are placed in them (Fig. 2c). Then new questions are made for the child nodes, and the samples are divided further (Fig. 2d). This results in a tree, in which non-leaf nodes contain questions about the samples, and

Sample	x	y	Class
A	0.2	1	1
B	0.9	1	1
C	0.8	2	2
D	0.4	2	1
E	0.5	2	1
F	0.8	3	1
G	0.9	2	2

(a)



(b) (c) (d)

Figure 2: An example of a decision tree.

the leaf nodes contain samples which have same answers through the path from the root node to the leaf node. The questions should be selected so that the resulting nodes would be as *pure* as possible, i.e., containing only samples of one class. In the figure, the questions have only two possible answers (yes or no), which leads to a *binary decision tree*, but in principle, more general trees could be created. However, in this paper, only binary decision trees are considered.

Because finding a tree, which is globally optimal, is completely intractable problem [Breiman et al., 1984], a greedy optimisation method is normally used[1]. The usual way to construct the tree is to find a node and a question which results in as pure child nodes as possible. This procedure is repeated until the some constraint is achieved. Typically, the algorithms do not split nodes further if the number of samples in the node is too small, or if the improvement gained from the splitting of the node would be too small.

In [Gelfand et al., 1991], an iterative method for determining a good tree size was developed. First, the training data is divided in two subsets. The tree is grown using the first set, until each leaf node contains only members of a one class, or has fewer samples than specified. This overly large, and most likely over-fitted tree, is then pruned using the second data set. Minimising the misclassification rate of the second data set, the optimal pruned tree is selected among all possibilities. Then the roles of the two data sets is changed, and the tree is grown and pruned again. This process is repeated until the tree converges.

3 Trees in acoustic modelling

3.1 Tying models or states

The use of decision trees in acoustic modelling was introduced in [Bahl et al., 1991]. Compared to the previous work done on clustering triphones, the decision tree based method allowed also the use of longer contexts, and actually five preceding and five following phonemes were studied in the paper. To select the actual HMM used for a triphone, questions about the neighboring phonemes were used to create a decision tree. Questions were selected among 1300 fixed question, and each of them addressed a single neighboring phoneme, for example, "is the previous phoneme a vowel?" or "is the next phoneme p, k or t?" The performance of the system was tested on a continuous speech recognition task with a vocabulary of 5000 words, and the decision trees decreased the error rate from 9.2% to 5.3%.

Instead of creating a decision tree for each phoneme, and using the tree to select the HMM for each context, [Young et al., 1994] created a tree for each state position of each phoneme, and used to tree to decide which *states* should share parameters. For example, triphones `t-i+s` and `d-i+s` could have separate first states, but common second and third state, as illustrated in Figure 3. This gives the decision tree algorithm much finer control over tying the parameters. The key result obtained in the experiments was that state-based clustering decreases the word-error rate 14% on average.

Also in [Odell, 1995], a decision tree is used to select HMMs for triphone contexts. In the thesis, a maximum likelihood based criterion is derived for selecting optimal questions for each node. Because it would be computationally too expensive to fit the acoustic models to the training speech data every time the tree is modified, several assumptions have been made. For example, it is assumed that the changes in the decision tree does not affect the state-frame alignment of the training data. Also, the transition probabilities are ignored.

[1] Actually, [Lazaridès et al., 1996] tried lookahead methods to obtain globally better trees, but surprisingly they found that greedy algorithm leads to better trees.

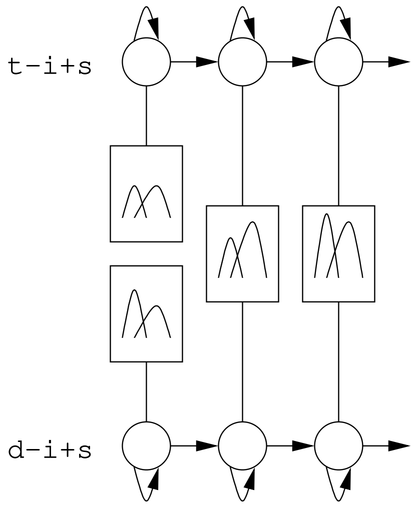


Figure 3: Example of two phoneme models `t-i+s` and `d-i+s`. The last two states have common parameters.

3.2 Additional questions

One of the biggest advantages of the decision trees is that they do not limit the scope of the questions in any way. This makes it possible to incorporate different sources of information in the decision process. In [Lazaridès et al., 1996] and [Gilles and Patrick, 1996], for example, questions about the state's position in the triphone was included in the tree. So, instead of making a separate tree for each state position of the phoneme, one tree was created for each phoneme, and tying between the state positions was also allowed. However, the advantage of this approach does not seem to be clear: [Lazaridès et al., 1996] report that, the questions about the state position did not improve the results, while [Gilles and Patrick, 1996] report that the questions improve error rates besides reducing the size of the model.

Another kind of questions were tried in [Fügen and Rogina, 2000]. In addition to the context and state positions, questions about the gender and dialect of the speaker, speaking rate, and signal-to-noise ratio were asked. Using the Janus speech recognition system, error rate reductions of 10% were achieved. They also report, that dynamic features like signal-to-noise ratio and speaking rate are a little problematic. In order to avoid reorganising the decoder-tree every time the features change, they shifted the questions about the dynamic features to the leaves of the tree, and used a lookup table to store references to acoustic models for each combination. With this technique, it was possible to achieve near real-time recognition.

3.3 Designing the questions

Most of the systems using decision trees to tie the parameters of the acoustic models, use a predefined set of questions. These questions are typically designed by an expert using prior knowledge on the problem, and refer to grammatical classes like *vowels* and *fricatives*. The problem with hand-crafted questions is, however, that the questions have to be redesigned, if the phoneme set changes. To avoid this kind of problems, an automatic method for generating the question set was developed in [Beulen and Ney, 1998].

The algorithm tries to find phoneme classes which then could be used in questions when creating the decision tree. Starting from triphone states grouped according to their central phoneme or diphone context, the groups are clustered using a bottom-up approach with a log-likelihood distance measure. The intermediate clusters during the process are stored and used as an initial question set. Finally, least important questions are discarded and questions about individual phonemes and word boundaries are added. Speech recognition experiments with this automatically generated question set showed that it can achieve the performance of a good hand-crafted question set.

4 Trees in language modelling

4.1 Language modelling

In the field of speech recognition, the most popular use of decision trees seems to have been in acoustic modelling, but trees have also been used in language modelling. And actually, before used for tying HMM parameters, decision trees were used to improve n-gram language models.

In speech recognition, the purpose of the language models is to assign probabilities to word sequences. Because in natural speech, some word sequences are very frequent and some word sequence, being meaningless, are very rare, language models can help speech recognition systems considerably. The most traditional language model, which also has been surprisingly effective, has been the n-gram model. Given a word sequence $W = w_1 w_2 \ldots w_n$, the n-gram model approximates the exact probability

$$P(W) = \prod_{i=1}^{n} P(w_i | w_1 \ldots w_i)$$

by taking into account only the $n - 1$ preceding words. For example, the 3-gram model gives the following probability to a word sequence W:

$$P(W) = P(w_1)P(w_2|w_1) \prod_{i=3}^{n} P(w_i|w_{i-2}w_{i-1}).$$

4.2 Tree-based language models

In [Bahl et al., 1989], a decision tree based language model was compared to a traditional 3-gram model. The words were grouped in grammatical classes such as days of the week, months, nouns, verbs, etc., and each word could belong to several classes. Then compound questions like “Is the preceding word an adjective and the one before that an article?”, or, “Is there any computer jargon in the last fifteen words?”, could be asked. The questions could refer up to 20 preceding words. The prediction performance of the models was tested, and lowest perplexity results were achieved using a combined trigram-tree model.

In [Heeman and Allen, 2001], more extra information was brought in language modelling. In fluent speech, people tend make intonational pauses, and corrections to the speech. Especially corrections degrade the performance of traditional language models that do not take them into account. In addition to the identity and part-of-speech tags of the words, the authors’ model introduced additional variables which indicated the range of corrections, and possible filler words such as “um” or “er”. All this information was used to create a decision tree based language model. The experiments gave encouraging results on detecting and correcting intonational boundaries and speech repairs, but speech recognition tests were not performed.

Semantic Classification Trees are another kind of specialised decision trees, which [Kuhn and De Mori, 1995] used as a building block for a speech understanding system. Instead of having questions, the nodes of the trees contain regular expressions made of words and possible gaps. These expressions can identify small key phrases, and when combined in a tree, a speech controlled ticket vending machine, for example, could decide whether the user asked to see the price of the tickets.

5 Summary

While not becoming a standard part of modern speech recognition systems, like hidden Markov models, decision trees have been used successfully to improve recognisers. Whether using long contexts in acoustic modelling or language modelling, the rapid increase in model parameters and lack of training samples cause easily over-fitting problems. Decision trees provide a convenient way to concentrate the model parameters in essential parts of the sample space. Their advantages are the possibility to combine different information sources, and the capability to handle missing data. The biggest drawback is perhaps that the questions, that are asked in the nodes of the trees, must often be designed manually, though automatic methods have been developed in some cases as described in Section 3.3.

References

[Bahl et al., 1989] Bahl, L. R., Brown, P. F., de Souza, P. V., and Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):1001–1008.

[Bahl et al., 1991] Bahl, L. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., and Picheny, M. A. (1991). Decision trees for phonological rules in continuous speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91)*, volume 1, pages 185–188.

[Beulen and Ney, 1998] Beulen, K. and Ney, H. (1998). Automatic question generation for decision tree based state tying. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, pages 805–808.

[Breiman et al., 1984] Breiman, L., Stone, C. J., Olshen, R. A., and Friedman, J. H. (1984). *Classification and regression trees.* Wadsworth, Inc.

[Fügen and Rogina, 2000] Fügen, C. and Rogina, I. (2000). Integrating dynamic speech modalities into context decision trees. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, volume 3, pages 1277–1280.

[Gelfand et al., 1991] Gelfand, S. B., Ravishankar, C. S., and Delp, E. J. (1991). An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):163–174.

[Gilles and Patrick, 1996] Gilles, B. and Patrick, K. (1996). Optimal tying of HMM mixture densities using decision trees. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 350–353.

[Heeman and Allen, 2001] Heeman, P. A. and Allen, J. F. (2001). Improving robustness by modeling spontaneous speech events. In Junqua, J.-C. and van Noord, G., editors, *Robustness in Language and Speech Processing*, chapter 5. Kluwer Academic Publishers.

[Kuhn and De Mori, 1995] Kuhn, R. and De Mori, R. (1995). The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):449–460.

[Lazaridès et al., 1996] Lazaridès, A., Normandin, Y., and Kuhn, R. (1996). Improving decision trees for acoustic modeling. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 2, pages 1053–1056.

[Odell, 1995] Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge.

[Young et al., 1994] Young, S. J., Odell, J. J., and Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 307–312.