Assignment 1 - Copy models for data compression

Algorithmic Information Theory (2022/23)
Universidade de Aveiro

Martinho Tavares, 98262, martinho.tavares@ua.pt
Nuno Cunha, 98124, nunocunha@ua.pt
Pedro Lima, 97860, p.lima@ua.pt

March 29, 2023

1 Introduction

One of the goals of this assignment is to implement a copy model for data compression, which is obtained through the exploration of self similarities. A copy model is only one of the ways to address the problem, and its idea is that certain data sources can be viewed as a sequence of repetitive patterns, where parts of the data have been replicated in the past, although some modifications are possible. The model consists in predicting the next symbol based on previous ones observed in the past, by keeping a pointer referring to the symbol being copied, as well as other information.

The second goal of the assignment is to implement a text generator based on the copy model, which is a way to generate new text based on a given one. The text generator receives a text as input to train the model, and then follows a similar approach to the one used in the copy model, but instead of predicting the next symbol, it uses the probability distribution to generate a random symbol based on these probabilities.

In this report we will first present how the work was organized, in the 2 section. Then, we will present the 3, how we implemented it and the results obtained, which we will compare by calculating the entropy of different text examples, like `chry.txt` given by the teachers, with different parameters we defined for the model. The next section is dedicated to the 4, where we will present the implementation and the results obtained for the text generator, using different texts as input for training of the generator, as well as diferrent starting points for the generation of text. Finally, we will conclude the report in the 5 section.

2 Work organization

3 Copy model

The copy model behaves in a straightforward way. At each symbol in the file, the program verifies if the pattern of k symbols with the current symbol being the last has already occurred in the past. That is, at position x_n, the pattern within positions x_{n-k+1} and x_n is evaluated to see if it appeared in the past.

To do that, each of these k-sized patterns at every position is saved in a hash table built incrementally while the file is being processed. The keys are the $k$mers/patterns, and the values are arrays of positions in the file where those $k$mers occurred in the past, which we call pointers. When we need to choose a pattern in the past, we look through the array of pointers for that pattern for a specific occurrence of that pattern.

Whenever the pattern is still not present in the hash table, we try to randomly guess which symbol should occur next since we can't perform a copy from the past. In these cases, the reported probabilities follow a fixed default distribution that is set beforehand.

When predictions can be made, a strategy is followed to choose the appropriate pointer for the current pattern. After a pointer is chosen, it is fixed and predictions are made sequentially from that point on. That is, if at position x_n we have the copy pointer x_m, then we will predict the character at x_{n+1} to be equal to the character at x_{m+1}, the character at x_{n+2} to be equal to the character at x_{m+2}, and so on. The predictions aren't completely certain; we assume a probability of making a correct prediction, which varies depending on how well the prediction going.

This process is repeated until a threshold is met, which is based on the probability of making a correct prediction at the current position. When the threshold is met, we stop predicting, and keep evaluating the file's content as normal. If the current $k$mer has already occurred in the past, then we attempt a new chain of predictions. Otherwise, we perform guesses.

We need to be careful regarding the beginning of the file. For the first $k-1$ symbols, it is not possible to generate a $k$mer, but we should still be able to perform predictions starting on these $k-1$ symbols. We followed a simple strategy, which involves extending the file content backwards, generating a past of size k. The generated past is simple, it is composed of k repetitions of a single symbol of the alphabet. This symbol is the most frequent of the entire alphabet.

In order to evaluate whether the copy model can provide acceptable results, we can use a baseline below which we expect the model to report the file's entropy. We decided to use, as a baseline, the entropy considering each symbol's relative frequency in the entire file, which is given by:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

With this value in mind, we evaluated the model as a whole with different values for its parameters, on different files. The files chosen for testing are present in the repository[1], and they have the following baselines:

- `chry.txt`: ...
- ...: ...

Throughout this section, the different program parameters are detailed, and their effect on the model's performance, both in terms of execution time and quality of the results, is studied.

For the results regarding the information content of the symbols throughout the `chry.txt` file, the data was transformed with a low-pass filter using a moving average, since there are too many data points.

3.1 Pattern size

When choosing a pointer in the past from which to start copying, we need to look for an occurrence of the same k-sized pattern as the one we are currently on. Thus, k is one of the parameters that affects program performance, where k is a positive integer.

On one hand, a lower value of k means there will be more occurrences of each pattern, since there will be lesser possible k-sized patterns, which will result in more oportunities to copy from the past. On the other, a larger k will cause less occurrences for each pattern, which results in more scarce predictions, and thus more overall guesses.

For a given k, the number of possible patterns M is equal to the number of permutations of the symbols of alphabet A: $M = |A|^k$. The results on changing this value will be studied taking this exponential growth into account.

3.1.1 Results

...

3.2 Smoothing parameter alpha

At each coding step, the probability of correctly predicting the current character is given by the formula:

$$P(\text{hit}) = \frac{N_h + \alpha}{N_h + N_f + 2\alpha}$$

where N_h and N_f are respectively the number of hits and misses, and α is the smoothing parameter. The smoothing parameter is used to avoid the model having a probability of 0 when there are no hits or misses, which will happen when no prediction has been made yet, and also avoid assuming with complete

[1] `https://github.com/NMPC27/TAI-G7-Lab1`

certainty (probability of 1) that the prediction will be correct when no fails were made yet, which will result in infinite information in case the prediction fails.

If α is lower, we expect the reported probabilites to be more "certain" with a greater possibility of reaching high values. If α is greater, the smoothing effect should be more pronounced, producing probability values that hardly increase, which should result in relatively less information when predictions fail but also lesser returns on the correct predictions.

This trade-off between α and the reported information, especially regarding the information spikes that may occur, will be studied in the results.

3.2.1 Results

...

3.3 Base probability distribution

For each symbol in the file, a probability distribution has to be generated according to the prediction probability. In case a given prediction fails, we need to distribute the complement of the probability of a correct prediction $1 - P(\text{hit})$ among the other symbols of the alphabet, except the one that we predicted would occur (which has probability $P(\text{hit})$).

The choice of distribution also influences the guesses that are made. When a prediction can't be made, the model assumes a distribution for the entirety of the alphabet, which is the same probability distribution as the use used when distributing the complement of the prediction probability.

We developed two simple approaches to distributing the probability: a uniform distribution and a frequency distribution.

The uniform distribution assigns to each symbol a probability equal to:

$$\frac{1 - P(\text{hit})}{|A| - 1}$$

On the other hand, the frequency distribution assigns to each symbol $a \in A$ a probability equal to:

$$(1 - P(\text{hit})) \times p(a)$$

where $p(a)$ is the probability of a in the whole file.

Both distributions are compared especially in their effect on the overall reported information, when k is large and thus more guesses are made, as well as when α is large and the prediction fails should report. We assume that, since the frequency distribution models the overall distribution of the symbols, it should provide better results. We should note, however, that the file's content may have locality, i.e. select portions in the file where the symbol distribution is skewed. Therefore, we don't disregard the possibility that the assumption of the distribution being maintained throughout the file can hurt the model performance as well.

3.3.1 Results

...

3.4 Copy pointer repositioning strategy

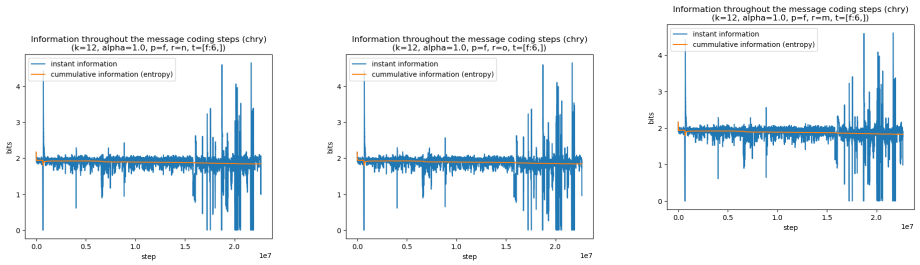
Different strategies can be followed when choosing a copy pointer from which to start copying. We developed 3 approaches, 2 of which being fairly simplistic:

- **Oldest pointer first**: the chosen copy pointer is the oldest possible. If the current copy pointer is repositioned because the threshold was met, then that pointer will be completely discarded and the next oldest pointer is used. This is the most simple approach, which assumes similar patterns are very far in the file;
- **Newest pointer first**: the chosen copy pointer is always the newest, which assumes that similar patterns are very close in the file;
- **Most common prediction**: when a pointer has to be chosen, the predictions of all registered pointers are evaluated, and the pointer (or one of them) which presented longest and most common prediction is chosen. A more detailed description of this strategy is presented below.
- **Most common prediction bounded**: the approach is exactly the same as the **most common prediction**, with the only difference being that the registered pointers for each pattern are stored in a circular array with limited size.

The **most common prediction** strategy chooses the copy pointer in successive iterations. Initially, the next immediate character is evaluated for every registered pointer. The pointers are grouped according to that character that they predict, and the largest group is chosen, i.e. the pointers that reported the most common prediction are chosen. Afterwards, the second next character is evaluated for every pointer in that group, and then the pointers are grouped in the same manner again, with the next group being the largest at this stage. This process is repeated until all pointers at the current group report different predictions, at which point the newest pointer is chosen.

In order to obtain the most common prediction at each iteration, the Boyer-Moore majority vote algorithm was used [1]. This is a very space-efficient algorithm (complexity $O(1)$) that guarantees the majority value will be returned. However, if a majority value doesn't actually exist then the algorithm may erroneously report a value as the majority, if we only do one pass. Since in this case we don't have a specific disambiguation strategy when the pointers don't agree, the choice is arbitrary and thus this is not a problem.

The bounded version of **most common prediction** uses a circular array for the set of registered pointers. The idea is for the array of registered pointers to contain only the N most recent pointers, with N being the array size. This can not only take into account the locality that the file's content may have, i.e.



(a) Oldest pointer first. (b) Newest pointer first. (c) Most common prediction.

Figure 1: Information of the current symbol and entropy as the file is processed, varying the pointer reposition strategy.

the repeated patterns being close to each other, but it also bounds the array size to a fixed value, which helps in performance since the reposition algorithm previously described has linear time complexity $O(N)$, when the other strategies have only $O(1)$ complexity.

3.4.1 Results

3.5 Copy pointer reposition threshold

A copy pointer is used until the predictions are deemed to not be worthwhile. After that point, the copy pointer is repositioned taking the current pattern into account. The decision of when to reposition the copy pointer can be done using different approaches.

We developed 3 strategies, all of which only factor in the prediction probability over time:

- **Static**: a static probability threshold value. If the prediction probability falls below this value, then the pointer is repositioned.
- **Successive fails**: if the number of successive prediction misses is larger than a given value, then the pointer is repositioned. In this case, the "successive fails" are evaluated using a non-negative counter of misses that is updated as predictions are made; whenever there is a miss, the counter increases, and decreases otherwise.
- **Rate of change**: if the absolute difference between the current prediction probability and the previous prediction probability is below a threshold then the pointer is repositioned.

Each strategy has a threshold value associated with them, and they are passed as parameters along with the specified strategy. These thresholds can also be combined together, indicating whether the pointer should be repositioned when at least one of the strategies reports that their threshold was surpassed.

3.5.1 Results

The results were initially evaluated for each strategy in isolation. After determining the best performing threshold values for each approach, they were combined among themselves, testing the 4 possible combinations.

...

3.6 Verbose output

As a sidenote, the model offers the possibility of outputting the probability distributions and predicitons at each step, or of presenting the overall progress of processing the input file.

For the modes outputting the probabilities, the output can be in human-readable format, presenting whether the reported probabilites were from a guess or a prediction hit/miss in colored output. Alternatively, the output can be presented as comma-separated values with less extraneous characters in machine-readable format, which is useful for data analysis afterwards.

4 Text generator

The Generator module makes use of the functionalities offered by the CPM module. This module uses a Context Model calculated from a received text to calculate the next characters based on a context phrase.

It contains two main methods, the `firstPass()`, that reads the training file and makes an initial pass through the text to calculate the frequencies of the characters. This method also makes an context model based on the training file, and the size of the context is defined by the input string.

The `cpm_gen()` method uses the context model in the first pass to generate the text. When we try to generate the next symbol, we read the current context, which is the size of the input string and we check the context model for that context. If we have that context in the model, we get the list of possible next symbols, and their counts. Then, we generate a random value between 0 and the sum of the counts, and we check where that value lands between all the ranges of counts of events of that context. The symbol that is in the range of the random value is the next symbol of the text. If we don't have that context in the model, we get the list of all possible symbols in the alphabet, and their distribution, and we generate a random value between 0 and the sum of the distributions, and we check where that value lands between all the ranges of distributions of events of that context. The symbol that is in the range of the random value is the next symbol of the text.

We also have 2 other input parameters that modify the behavior of the generator. The first one is allowing the model to train itself, while generating the text. This is done by updating the context model with the generated text and the distibution of symbols in the alphabet. The second parameter is to allow all lowercase characters to be generated, this is done by converting all the characters to lowercase while reading the training file. This parameter provides

a better context model without needing a big training file, but it also makes the text generated to be all lowercase.

4.1 Results

Results for the text generator without -l or -t parameters:

Listing 1: Generator output for train file "othello.txt" with initial string "T" and without -l or -t parameters
Num guesses: 0

```
Houllo m  oue he plled th So taby t want,    ke tirrad:
```

Listing 2: Generator output for train file "othello.txt" with initial string "Th" and without -l or -t parameters
Num guesses: 0

```
I the my And him, Iager'd and th
```

Listing 3: Generator output for train file "othello.txt" with initial string "The" and without -l or -t parameters
Num guesses: 0

```
Canitiall with Office. But that Cassio's see spokes.
```

Listing 4: Generator output for train file "othello.txt" with initial string "The " and without -l or -t parameters
Num guesses: 0

```
Othello's Let to obey too much a round it, I thinke.
```

Listing 5: Generator output for train file "othello.txt" with initial string "The t" and without -l or -t parameters
Num guesses: 0

```
I would them. And yet went busie must be Iago. Liue Rodo.
With this?
```

Listing 6: Generator output for train file "othello.txt" with initial string "The te" and without -l or -t parameters
Num guesses: 165

```
And marke how he comes to make a word or two before
apprehend her faire Island,

Dc dHmr i aeihe, ooeelde smgimfpwi ?
```

Listing 7: Generator output for train file "othello.txt" with initial string "The tes" and without -l or -t parameters
Num guesses: 8 311

```
Do'st thou hast no name the Instruments summon to supper
time

hatw,es eegpa rdlia vlnanrias in o ry ESowmn b syess g
lant nero ue
```

Listing 8: Generator output for train file "othello.txt" with initial string "The test" and without -l or -t parameters
Num guesses: 10 000

```
one en,eLmosiuhnunrgtlveecm.n .aulebalom hdrd
```

Results for the text generator with -l parameter:

Listing 1: Generator output for train file "othello.txt" with initial string "T" and with -l parameter
Num guesses: 0

```
e, t t fllfod is pe ale mout thort g ld da meli s area
vperd
```

Listing 2: Generator output for train file "othello.txt" with initial string "Th" and with -l parameter
Num guesses: 0

```
dosto th torieuers: and fat crythers so yet so:
```

Listing 3: Generator output for train file "othello.txt" with initial string "The" and with -l parameter
Num guesses: 0

```
and timentena some rod. if that now he drunke more ete.
lord?
```

Listing 4: Generator output for train file "othello.txt" with initial string "The " and with -l parameter
Num guesses: 0

```
my wife: i would nothinkings. what i will men othello,
take thy awl'd thing.
```

Listing 5: Generator output for train file "othello.txt" with initial string "The t" and with -l parameter
Num guesses: 0

```
rodo. patience. let's go safely by breake no iago. 'faith
one cases,
```

Listing 6: Generator output for train file "othello.txt" with initial string "The te" and with -l parameter
Num guesses: 0

```
why stand i loue be now to dyet my lippes: is't you cassio
be receiu'd her father,
```

Listing 7: Generator output for train file "othello.txt" with initial string "The tes" and with -l parameter
Num guesses: 7 899

```
that i can discourse. which they iumpe not one, that
sweete othello?

wuo og,ahwolseh.eu afet hfmgath saegnnr hisdsdhaaiin
```

Listing 8: Generator output for train file "othello.txt" with initial string "The test" and with -l parameter
Num guesses: 9 229

```
mus. i marry. what? a custome: but in a man that latest,
which not enriches him,
```

```
tt oitdi ?w,ehyas a lw n rhge hayo
```

Results for the text generator with -t parameter:

Listing 1: Generator output for train file "othello.txt" with initial string "T" and with -t parameter
Num guesses: 0

```
So Ieread atle m bo. wathin o. Lino Whe Ansois. tes sh.
```

Listing 2: Generator output for train file "othello.txt" with initial string "Th" and with -t parameter
Num guesses: 0

```
We sioneues, makendinget want hon and, youry her.
```

Listing 3: Generator output for train file "othello.txt" with initial string "The" and with -t parameter
Num guesses: 0

```
Iago. Why Minuation, that pray you know is Tempt is liues of it same
```

Listing 4: Generator output for train file "othello.txt" with initial string "The " and with -t parameter
Num guesses: 0

```
Exit Cas. I will you go with loue alread Rod. What breaken of Maine:
```

Listing 5: Generator output for train file "othello.txt" with initial string "The t" and with -t parameter
Num guesses: 0

```
Othello: heereafter more next night, or his faire him at your Daughter
```

Listing 6: Generator output for train file "othello.txt" with initial string "The te" and with -t parameter
Num guesses: 34

```
Des. It is now a huge Eclipse

a d sia r io a proper man: Oh, my face
```

Listing 7: Generator output for train file "othello.txt" with initial string "The tes" and with -t parameter
Num guesses: 6 533

```
How he vpbraides Iago, Attendants, with leade direction.

mpo gu hsees w?u a sfeesrso  hsnesatlAghhoif
```

Listing 8: Generator output for train file "othello.txt" with initial string "The test" and with -t parameter
Num guesses: 10 000

```
s:cieBgt kofhb  afgeerarHlyuhm.rlwuy.ett u tihlA mmr'
inaaDnr nuI'oWanulo
```

5 Conclusion

6 References

References

[1] R. S. Boyer and J. S. Moore, *MJRTY—A Fast Majority Vote Algorithm*. Dordrecht: Springer Netherlands, 1991, pp. 105–117. [Online]. Available: https://doi.org/10.1007/978-94-011-3488-0_5