Faster universal modeling for two source classes

Faster Universal Modeling for Two Source Classes

Ali Nowbakht* and Frans Willems

Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract. The Universal Modeling algorithms proposed in [2] for two general classes of finite-context sources are reviewed. The above methods were constructed by viewing a model structure as a partition of the context space and realizing that a partition can be reached through successive splits. Here we start by constructing recursive counting algorithms to count all models belonging to the two classes and use the algorithms to perform the Bayesian Mixture. The resulting methods lead to computationally more efficient Universal Modeling algorithms.

1 Introduction

We review the Universal Modeling algorithms proposed in [2] for two finite-context source classes : Class-I and Class-II. These algorithms were developed in the framework of a generalization of the Context-Tree Weighting (CTW) algorithm [1] and perform a recursive weighting of all models. A close look into the workings of the methods for Class-I and Class-II reveals that some of the models are being counted in repeatedly. This results in excessive storage and a higher computational complexity than strictly needed. Using an approach based on counting algorithms inside the original methods we can remove these repetitions and therefore reduce the complexity without sacrificing the performance.

2 Universal Source Coding Overview

The purpose of Source Coding is to represent sequences in the most compact way. This representation is called a (source) code and each (binary) sequence $x_1^T = x_1, \ldots, x_T$ of length T is represented by a (binary) codeword $c(x_1^T)$ of length $L(x_1^T)$. According to the concept of entropy of Shannon the ideal codeword length is related to the probability of the sequence $P(x_1^T)$ by the following expression $L_{id}(x_1^T) = -\log_2 P(x_1^T)$ bits. The probability of a sequence depends on the characteristics of the information source which generated it. In the Universal Source Coding setting the characteristics of the information source are unknown and therefore the probability $P(x_1^T)$ has to be estimated from the sequence x_1^T and any other available information.

* Supported by Technologiestichting STW under project EEL4643.

Finite-Context Sources These sources are characterized by the fact that their current state (parameter) is determined by the current context information through some function $\beta(\cdot)$. For each sequence symbol x_t there is a context symbol u_t available. Therefore we can express the instantaneous probability as $P(X_t = 1) = 1 - P(X_t = 0) = \theta_{\beta(u_t)}$ for $t = 1, \ldots, T$ where $u_t \in \mathcal{U} = \{1, \ldots, U\}$ the *context space* and $\theta_k \in [0, 1]$ is a parameter. The probability of the whole sequence would be $P(X_1^T = x_1^T) = \prod_{t=1}^{T} P(X_t = x_t)$.

Structure and Parameters The Source Model consists of two parts, namely the Model Structure (determined by the function $\beta(\cdot)$) and the Source Parameters specified by the Parameter Vector $\Theta = \{\theta_k, k = 1, \ldots, K\}$ (K is the number of parameters here). The Structure specifies which groups of contexts correspond to the same state (parameter), these are called Context-Sets. The structure can be seen as a partition of the context space into disjoint subsets. The parameters define the probability distribution for every state. A Source Class is a collection of structures that satisfy some restrictions on the allowed context-sets i.e. only some groups of contexts can correspond to the same state.

Universal Model - Bayesian Mixture A Universal Model is a probability distribution that fits any source model. A conceptually straightforward way to construct such a universal model is to perform the Bayesian Mixture.

$$P_c(x_1^T) = \sum_{M \in \mathcal{M}} P_{\mathcal{M}}(M) P(x_1^T \mid M) \tag{1}$$

In the above expression $P_c(x_1^T)$ is the universal probability assigned to string x_1^T. It is constructed by weighting (averaging) $P(x_1^T \mid M)$, the probabilities assigned by each structure M from source class $\mathcal{M}$, with the *a-priori* probability of that structure $P_{\mathcal{M}}(M)$. $P(x_1^T \mid M) = \prod_{i=1}^{K} P_e(\mathcal{S}_i)$, where $P_e(\mathcal{S})$ is an estimate for the probability of the subsequence corresponding to all symbols which where generated with context $u \in \mathcal{S} \subseteq \mathcal{U}$. Here model M partitions $\mathcal{U}$ into K cells $\mathcal{S}_i$.

3 Universal Modeling for Class-I

Class-I is the most general source class one can think of since it makes no restrictions whatsoever on the composition of the context-sets.

Context-Sets and Structures If the size of the context space is n there are $\binom{n}{s}$ possible context-sets of size s. Hence, in total there are $\sum_{s=1}^{n} \binom{n}{s} = 2^n - 1$ different subsets. The number of different model structures is the number of distinct partitions of the context space into disjoint subsets. In general, $N(n, p)$ the number of partitions of a context space of cardinality n into p subsets can be expressed as $N(n, p) = \frac{M(n,p)}{p!}$ where $M(n, p)$ defines the number of partitions into p *labeled* subsets and is defined recursively $M(n, p) = p^n - \sum_{i=1}^{p-1} \binom{p}{i} M(n, i)$. The total number of structures in Class-I is thus $\sum_{i=1}^{n} N(n, i)$.

3.1 The Arbitrary Splitting Method

As should be obvious from the preceding section, it is infeasible to calculate the Bayesian Mixture (1) by summing all models one by one. Therefore Willems *et al.* proposed the *Arbitrary Splitting* (AS) method in [2] as an alternative. We describe briefly the AS method. For each of the possible context-sets $\mathcal{D}$ a record is hold which keeps two probabilities : the *Estimated Probability* $P_e(\mathcal{D})$ and the *Weighted Probability* $P_w(\mathcal{D})$.

$P_e(\mathcal{D})$ is an estimate for the probability of the subsequence corresponding to all symbols which where generated with context $u \in \mathcal{D}$.

The weighted probability $P_w(\mathcal{D})$ is defined as the uniform weighting of the estimated probability and the weighted probabilities of all substructures which result by splitting $\mathcal{D}$ (into two subsets). The set $\Pi(\mathcal{D})$ of all possible splits of context-set $\mathcal{D}$ is defined in the following manner

$$\Pi(\mathcal{D}) = \{(\mathcal{S}_1, \mathcal{S}_2) : \mathcal{S}_1 \neq \emptyset, \mathcal{S}_1 \neq \mathcal{D}, \mathcal{S}_2 = \mathcal{D} \setminus \mathcal{S}_1\}$$

The weighted probability is expressed as

$$P_w(\mathcal{D}) = \frac{P_e(\mathcal{D}) + \sum_{(\mathcal{S}_1, \mathcal{S}_2) \in \Pi(\mathcal{D})} P_w(\mathcal{S}_1) P_w(\mathcal{S}_2)}{2^{|\mathcal{D}|-1}} \tag{2}$$

The Universal Model is defined as $P_c(x_1^n) = P_w(\mathcal{U})$ and includes all models in a recursive way.

Computational Complexity For a certain subset $\mathcal{D}$ we need according to expression (2) $2^{|\mathcal{D}|-1} - 1$ additions and $2^{|\mathcal{D}|-1}$ multiplications. Note that we are just counting the operations needed to calculate $P_w(\mathcal{D})$ when all involved terms are already available. Summing up over all subsets will give the total complexity. Remember that for a context space of cardinality n there are $\binom{n}{s}$ subsets of size s. The total number of additions is therefore

$$N_{add}^{AS}(n) = \sum_{s=1}^{n} \binom{n}{s} (2^{s-1} - 1) = \frac{3^n - 2^{n+1} + 1}{2}$$

and the total number of multiplications

$$N_{mul}^{AS}(n) = \sum_{s=1}^{n} \binom{n}{s} 2^{s-1} = \frac{3^n - 1}{2}$$

Model Multiplicity There are many ways to arrive at a partition by means of successive splitting. Consider a model with p parameters i.e. a partition of the context space into p cells. We can easily write down the following recursive formula for $\mu(p)$ the number of different ways we can arrive at this particular model

$$\mu(p) = \sum_{i=1}^{\frac{p}{2}-1} \binom{p}{i} \mu(i)\mu(p-i) + \frac{1}{2}\binom{p}{\frac{p}{2}} \mu^2(\tfrac{p}{2}) \qquad \text{for } p \text{ even}$$

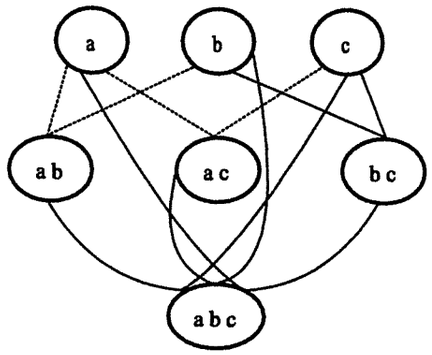

Fig. 1. All Context-Sets for Class-I with $\mathcal{U} = \{a, b, c\}$. Splits are shown for new method (only filled lines) and AS method (all lines).

$$\mu(p) = \sum_{i=1}^{\frac{p-1}{2}} \binom{p}{i} \mu(i)\mu(p-i) \qquad \text{for } p \text{ odd}$$

with $\mu(1) = 1$.

Example 1. In figure 1 the $\mu(3) = 3$ successions of splits leading to structure $\mathcal{S}_1 = \{a\}$ $\mathcal{S}_2 = \{b\}$ $\mathcal{S}_3 = \{c\}$ can be appreciated.

3.2 E1 : A New Universal Modeling Method

From the preceding section it should be clear that for the AS method it holds that the number of times a structure is included in the Universal Model explodes with its number of parameters. This observation motivates the search for methods which perform the Bayesian Mixture without repeating models in the sum and hence saving arithmetic operations. We start by introducing an algorithm to count all models in Class-I. Let $a(n)$ be the number of models in Class-I for a context space of size n. Let us begin counting all models by selecting an arbitrary context $x \in \mathcal{U}$, and consider all models where x forms a context-set on his own. Since there are $n-1$ contexts left, there will be $a(n-1)$ such models.

So far we have only considered the models with x forming a subset. Let us now add all models where x is joined by one of the other $n-1$ contexts, say y. Now $\{x, y\}$ form a context-set and there are thus $n-2$ contexts left. Therefore there are $(n-1) \cdot a(n-2)$ such models. By continuing in this way we can write a recursive formula for $a(n)$ namely

$$a(n) = \sum_{i=0}^{n-1} \binom{n-1}{i} a(n-1-i)$$

with $a(0) = 1$.

Obviously it must be true that $a(n) = \sum_{p=1}^{n} N(n,p)$. The above algorithm can be used to perform the Bayesian Mixture. First of all we define a set of splits of a generic context-set $\mathcal{D} \subseteq \mathcal{U}$. This set is formed by all splits used in the above algorithm when applied on $\mathcal{D}$. Consider an arbitrary element $x \in \mathcal{D}$. We define

$$\Omega(\mathcal{D}) = \{(\mathcal{S}_1, \mathcal{S}_2) : \mathcal{S}_1 \neq \emptyset, \mathcal{S}_2 = \mathcal{D} \setminus \mathcal{S}_1, x \in \mathcal{S}_1\} \tag{3}$$

$\Omega(\mathcal{D})$ contains all possible splits, the same as $\Pi(\mathcal{D})$, the difference is that here we ask that the sets containing x are called $\mathcal{S}_1$ or what is the same that all $\mathcal{S}_1$'s must have a common element.

Note that $\Omega(\mathcal{D})$ includes the void split $(\mathcal{D}, \emptyset)$ which $\Pi(\mathcal{D})$ does not, and therefore the number of splits is $|\Omega(\mathcal{D})| = 2^{|\mathcal{D}|-1}$.

The new method works in the following way. Again we have records holding the estimated probability $P_e(\mathcal{D})$ for all possible context-sets in Class-I. But now instead of having a weighted probability attached to all $(2^n - 1)$ context-sets we only need it for the context-sets which will be further split (the $2^{n-1} - 1$ sets called $\mathcal{S}_2$ in (3)). Only for these context-sets we define $P_f(\mathcal{D})$ probabilities

$$P_f(\mathcal{D}) = \frac{\sum_{(\mathcal{S}_1,\mathcal{S}_2)\in\Omega(\mathcal{D})} P_e(\mathcal{S}_1) P_f(\mathcal{S}_2)}{2^{|\mathcal{D}|-1}}$$

where $P_f(\emptyset) = 1$. This reduces the storage need for keeping the weighted probabilities to the half.

We define the Universal Model as $P_c(x_1^n) = P_f(\mathcal{U})$ and now each model is included only once in the sum since there is only one possibility to arrive at a partition through successive splits.

Example 2. The new method results in removing the dashed splits in figure 1.

Computational Complexity For a certain subset $\mathcal{D}$ we need according to the above expression $2^{|\mathcal{D}|-1} - 1$ additions and $2^{|\mathcal{D}|-1}$ multiplications. Summing up over all subsets for which a weighted probability is necessary will give the total complexity. Note that for a context space of cardinality n there are only $\binom{n-1}{s}$ context-sets of size s which have a $P_f(\cdot)$ attached. The total number of additions is therefore

$$N_{add}^{E1}(n) = \sum_{s=1}^{n-1} \binom{n-1}{s} (2^{s-1} - 1) = N_{add}^{AS}(n-1)$$

and the total number of multiplications

$$N_{mul}^{E1}(n) = \sum_{s=1}^{n-1} \binom{n-1}{s} 2^{s-1} = N_{mul}^{AS}(n-1)$$

In summary, this new approach increases the speed with respect to the AS method by a constant factor and reduces the storage need for keeping the weighted probabilities to the half.

4 Universal Modeling for Class-II

Class-II is defined by first considering a lexicographical ordering on the context space $\mathcal{U}$. Since we have defined $\mathcal{U} \doteq \{1, \ldots, n\}$ the lexicographical ordering is the usual ordering of the natural numbers. The only allowed partitions of the context space (because of the ordering it is now a line) are those which divide it into intervals, each forming a context-set. We introduce the following notation for specifying context-sets $(i \rightarrow j) \doteq \{u \in \mathcal{U} : i \leq u \leq j\}$ where $i, j \in \mathcal{U}$ and $j \geq i$.

Context-Sets and Structures Note that for a context space of size n there are $\frac{n(n+1)}{2}$ possible context-sets and that there are $\binom{n-1}{p-1}$ different structures having p parameters. This means that in total there are $\sum_{p=1}^{n} \binom{n-1}{p-1} = 2^{n-1}$ possible structures in Class-II.

4.1 The Lexicographical Splitting Algorithm

Although Class-II is a small class compared to Class-I it still includes an exponential number of structures making the brute-force approach to calculating the Bayesian Mixture infeasible. Therefore Willems *et. al* [2] proposed the *Lexicographical Splitting* (LS) algorithm. We describe the LS method briefly. For a context space of size n there is for each of the $\frac{n(n+1)}{2}$ possible context-sets $(i \rightarrow j)$ a record which keeps two probabilities : the *Estimated Probability* $P_e((i \rightarrow j))$ and the *Weighted Probability* $P_w((i \rightarrow j))$. $P_e((i \rightarrow j))$ is of course the same as in the methods for Class-I. The weighted probability $P_w((i \rightarrow j))$ is also defined in the same manner but its mathematical expression has to be adjusted to Class-II.

$$P_w((i \rightarrow j)) = \frac{P_e((i \rightarrow j)) + \sum_{k=i}^{j-1} P_w((i \rightarrow k)) \cdot P_w((k+1 \rightarrow j))}{j - i + 1} \quad \text{for } j > i \tag{4}$$

If $j = i$ we define $P_w((i \rightarrow i)) = P_e((i \rightarrow i))$. The Universal Model is $P_c(x_1^T) = P_w(\mathcal{U}) = P_w((1 \rightarrow n))$.

Computational Complexity We look now at the complexity of computing the weighted probability for a generic context-set of size d. Suppose that all weighted and estimated probabilities involved have been updated already, in that case we need $d - 1$ additions and d multiplications as can be seen from (4). Note that for a context space of size n there are $n - d + 1$ context-sets of size d.

$$N_{add}^{LS}(n) = \sum_{s=1}^{n} (n - s + 1) \cdot (s - 1) = \frac{(n-1)n(n+1)}{6}$$

$$N_{mul}^{LS}(n) = \sum_{s=1}^{n} (n - s + 1) \cdot s = \frac{n(n+1)(n+2)}{6}$$

The complexity is thus $O(n^3)$.

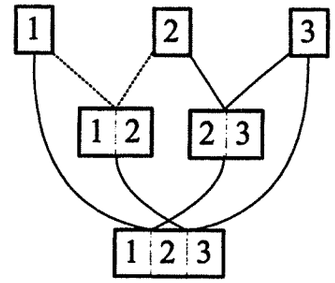

Fig. 2. All Context-Sets for Class-II with $\mathcal{U} = \{1, 2, 3\}$. Splits are shown for new method (only filled lines) and LS method (all lines).

Model Multiplicity This section is based on the observation that the LS algorithm arrives at some models through different splits. More precisely suppose a model having p parameters. The number of ways $\mu(p)$ we can arrive at a model with p parameters is given by the Catalan Numbers

$$\mu(p) = \sum_{i=1}^{p-1} \mu(i) \cdot \mu(p - i) = \frac{1}{p} \binom{2p-2}{p-1}$$

with $\mu(1) = 1$.

Example 3. In figure 2 the $\mu(3) = 2$ successions of splits leading to structure $\mathcal{S}_1 = \{1\}$ $\mathcal{S}_2 = \{2\}$ $\mathcal{S}_3 = \{3\}$ can be appreciated.

4.2 E2 : A New Universal Modeling Method for Class-II

As for Class-I we start by finding an counting algorithm for Class-II. Let $b(n)$ be the number of models in Class-II for a context space of size n. We can write an expression for all structures in function of the size of their first interval. $b(n) = \sum_{i=0}^{n-1} b(i)$ with $b(0) = 1$. Obviously $b(n) = 2^{n-1}$.

Therefore we define instead of the weighted probabilities, *Fast Weighting* probabilities $P_f(\cdot)$. Note that now we do not need to store a P_f in each of the $\frac{n(n+1)}{2}$ records corresponding to all context-sets $(i \rightarrow j)$ for $i = 1, \ldots, n$ and $j \geq i$. To calculate $P_f((1 \rightarrow n))$ we only need n P_f's, namely those corresponding to context-sets $(i \rightarrow n)$ for $i = 1, \ldots, n$.

$$P_f((i \rightarrow n)) = \frac{\sum_{k=i}^{n} P_e((i \rightarrow k)) \cdot P_f((k+1 \rightarrow n))}{n - i + 1}$$

where $P_f((n+1 \rightarrow n)) = 1$ by convention. The Universal Model is simply defined as $P_c(x_1^T) = P_f(1 \rightarrow n)$.

Example 4. The new method results in removing the dashed splits in figure 2.

Computational Complexity The only difference with respect to the analysis for the LS method is that now for a context space of size n there is only one context-set of size d which has a weighted probability attached.

$$N_{add}^{E2}(n) = \sum_{s=1}^{n} (s-1) = \frac{(n-1)n}{2}$$

$$N_{mul}^{E2}(n) = \sum_{s=1}^{n} s = \frac{n(n+1)}{2}$$

Which are of order $O(n^2)$.

In summary, this new method reduces the complexity from $O(n^3)$ to $O(n^2)$ and the storage need for keeping the weighted probabilities from $\frac{n(n+1)}{2}$ to n.

5 Conclusions

We have introduced new methods to perform the Bayesian Mixture for Class-I and Class-II. The difference to the earlier proposed methods of [2] can be best appreciated in the Model Multiplicity. In our methods each model can be reached through a unique succession of splits and therefore is included only once in the Universal Model. As we have shown, in the earlier methods this was not the case.

The new methods exhibit a lower computational complexity and storage need. In the case of Class-I the reduction is by a constant factor. For Class-II we go from $O(n^3)$ to $O(n^2)$ in complexity and from $O(n^2)$ to $O(n)$ in the storage need for the weighted probabilities. Here n represents the size of the context space.

6 Acknowledgment

The authors thank Henk van Tilborg and Stan Baggen for pointing them to the Catalan numbers.

References

1. F.M.J. Willems, Y.M. Shtarkov and Tj.J. Tjalkens, "The Context-Tree Weighting Method : Basic Properties," *IEEE Trans. Inform. Theory,* vol. 41, no. 3, pp. 653-664, 1995.
2. F.M.J. Willems, Y.M. Shtarkov and Tj.J. Tjalkens, "Context Weighting for General Finite-Context Sources," *IEEE Trans. Inform. Theory,* vol. 42, no. 5, pp. 1514-1520, 1996.