International Journal of Approximate Reasoning 103 (2018) 71–93

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Hierarchical two-part MDL code for multinomial distributions

Marc Boullé

Orange Labs, 22300 Lannion, France

ARTICLE INFO

Article history:
Received 12 June 2018
Received in revised form 3 September 2018
Accepted 7 September 2018
Available online 11 September 2018

Keywords:
Multinomial distribution
Minimum description length
Normalized maximum likelihood
Model selection
Density estimation
Information theory

ABSTRACT

We leverage the Minimum Description Length (MDL) principle as a model selection technique for multinomial distributions and suggest a two-part MDL code based on a hierarchical encoding of the multinomial parameters. We compare this code with the alternative Normalized Maximum Likelihood (NML) code and exhibit large regions of the parameter space where the hierarchical code dominates the NML one. We then present an application of the multinomial distribution to joint density estimation and show that the hierarchical code brings significant improvements.

1. Introduction

Industrial companies such as Orange, the main french telecommunication operator, store large amounts of data. They have to deal with many requests for data mining studies, in a wide diversity of application domains and tasks, structure and scale of data, constraints, resource or business requirements. To address these problems in an industrial context, Orange Labs has developed a data mining tool,[1] with the following requirements: generic, reliable, accurate, automatic, interpretable and scalable. This tool exploits models for conditional or joint density estimation in the univariate or multivariate cases, with either numerical or categorical variables [3], for feature selection and construction in the multi-tables context and for modeling [5]. All these models extensively use multinomial distributions as building blocks, and the inference process heavily relies on MDL model selection to meet the tool requirements. Enumerative codes have been used for years, being effective (they are both two-parts, one-part and NML codes) and very simple and efficient to compute at any scale. The objective of this paper is to study whether these codes can be improved in order to detect patterns with fewer instances, with the least possible computational overhead. In particular, we focus on the case of data sets with heavily unbalanced distributions, such as Zipf's law or Pareto distribution, which widely appears in many application domains such as linguistics, physics or economics [22,21].

Model selection is a key problem in statistics and data mining, and the MDL approaches [23] to model selection have been extensively studied in the literature [10], with successful applications in many practical problems. Simple models such as multinomial distributions are important because they are easy to analyze theoretically and useful in many applications. For example, the multinomial distribution has been used as a building block in more complex models, such as naive Bayes classifiers [19], Bayesian networks [30,12], decision trees [34] or coclustering models [3,11]. These models involve up to thousands of multinomial blocks, some of them with potentially very large numbers of occurrences and outcomes. For

E-mail address: marc.boulle@orange.com.

[1] This tool, named Khiops, is available as a shareware at www.khiops.com.

https://doi.org/10.1016/j.ijar.2018.09.002
0888-613X/

example, the text × word coclustering of the 20-newsgroup data set described in [3] exploits a main multinomial block with around two millions words (occurrences) distributed on 200,000 coclusters (outcomes). In [11], half a billion call detail records (occurrences) are distributed on one million coclusters (outcomes). These various and numerous applications critically rely on the use of effective and efficient MDL codes to get a robust and accurate summary of the data.

The MDL approaches come with several flavors, ranging from theoretical but not computable to practical but sub-optimal. Ideal MDL [33] relies on the Kolmogorov complexity, that is the ability of compressing data using a computer program. However, it suffers from large constants depending on the description method used and cannot be computed, not even approximated in the case of two-part codes [1]. Practical MDL leverages description methods that are less expressive than general-purpose computer languages. It has been employed to retrieve the best model given the data in case of families of parametrized statistical distributions. Crude MDL is a basic MDL approach with appealing simplicity. In two-part crude MDL, you just have to encode the model parameters and the data given the parameter, with a focus on the code length only. However, crude MDL suffers from arbitrary coding choices. Modern MDL relies on universal coding resulting in Refined MDL [10], with much stronger foundations and interesting theoretical properties. In particularly, the normalized maximum likelihood (NML) [25] provides a theoretically solid criterion based on a minimax regret strategy. The NML approach exploits a constant regret: all the distributions are treated on the same footing and the one that best fits the data is chosen. Interestingly, the enumerative two-part MDL code for multinomial models has a strong connection with the NML approach [6]. Despite its simplicity, this code is both a two-part and a one-part code, is optimal w.r.t. the NML approach and is parametrization invariant.

In this paper, we investigate on two-part codes for multinomial models based on a hierarchical encoding of the model parameters. Although they loose the appealing theoretical properties of the alternative NML code, they reach a better compression on large regions of the parameter space, namely in case of unbalanced multinomial distributions, with a negligible loss on the rest of the parameter space. We present an application of multinomial models to joint density estimation. We show that using the proposed hierarchical multinomial code significantly improves the quality of the retrieved models in the case of peaked densities, which closely relates to unbalanced distributions.

The rest of the paper is organized as follows. For self-containment reasons, Section 2 presents NML codes for the multinomial distribution. Section 3 introduces a hierarchical code for the multinomial distribution and compares it to alternative enumerative NML code. Section 4 presents an application of these codes to joint density estimation and analyzes the impact of the chosen code, from balanced to unbalanced data. Finally, Section 5 summarizes this paper.

2. NML codes for multinomial distribution

Let us consider the multinomial model with parameter $\theta = (\theta_1, \ldots, \theta_m)$, $\sum_{j=1}^{m} \theta_j = 1$, $\forall j, \theta_j > 0$, such that $P_\theta(X = j) = \theta_j$, in the case of m-ary sequences $x^n \in X^n$ of size n. For a given sequence x_n, $P_\theta(x_n) = \prod_{j=1}^{m} \theta_j^{n_j}$, where n_j is the number of occurrences of outcome j in sequence x^n.

2.1. Standard NML approach

Using universal coding, a grounded approach is proposed to evaluate the model complexity, based on the Shtarkov NML code [29], which provides strong theoretical guarantees [26].

It exploits the following NML distribution $\overline{P}_{nml}^{(n)}$ on X^n:

$$\overline{P}_{nml}^{(n)}(x^n) = \frac{P_{\widehat{\theta}(x^n)}(x^n)}{\sum_{y^n \in X^n} P_{\widehat{\theta}(y^n)}(y^n)} \tag{1}$$

where $\widehat{\theta}(x^n)$ is the model parameter that maximizes the likelihood of x^n.

The log of the denominator stands for the *parametric complexity* $COMP^{(n)}(\theta)$ of the model whereas the negative log of the numerator is the *stochastic complexity* of the data given the model. The sum of both terms provides the NML code. It is noteworthy that the NML code is a one-part rather than two-part code: data is encoded with the help of all the model hypotheses rather than the best hypothesis.

The parametric complexity of the NML universal model with respect to a k-parameter exponential family model is usually approximated by $\frac{k}{2} \log \frac{n}{2\pi}$ [10]. In the case of the multinomial distribution with $(m-1)$ free parameters, this gives $\frac{m-1}{2} \log \frac{n}{2\pi}$. A better approximation based on Rissanen's asymptotic expansion [25] is presented in [14]:

$$COMP_{nml}^{(n)}(\theta) = \frac{m-1}{2} \log \frac{n}{2\pi} + \log \frac{\pi^{m/2}}{\Gamma(m/2)} + o(1), \tag{2}$$

where $\Gamma(.)$ is the Euler gamma function. Still in [14], a sharper approximation based on Szpankowski's approximation [32] is presented. This last approximation, far more complex is very accurate w.r.t. n, with $o(\frac{1}{n^{3/2}})$ precision. We present below its first terms until $O(\frac{1}{\sqrt{n}})$.