Engineering Applications of Artificial Intelligence 63 (2017) 69–84

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Occlusion-based estimation of independent multinomial random variables using occurrence and sequential information ☆

B. John Oommen [a,c,*], Sang-Woon Kim [b]

[a] School of Computer Science, Carleton University, Ottawa, Canada K1S 5B6
[b] Dept. of Computer Engineering, Myongji University, Yongin 17058, Republic of Korea
[c] Department of Information and Communication Technology, University of Agder, Grimstad, Norway

ARTICLE INFO

Article history:
Received 19 February 2017
Received in revised form
7 April 2017
Accepted 5 May 2017
Available online 25 May 2017

Keywords:
Estimation using sequential information
Sequence Based Estimation
Estimation of multinomials
Fused Estimation Methods
Sequential information

ABSTRACT

This paper deals with the relatively new field of sequence-based estimation in which the goal is to estimate the parameters of a distribution by utilizing both the information in the observations and in their sequence of appearance. Traditionally, the Maximum Likelihood (ML) and Bayesian estimation paradigms work within the model that the data, from which the parameters are to be estimated, is known, and that it is treated as a set rather than as a sequence. The position that we take is that these methods ignore, and thus discard, valuable sequence -based information, and our intention is to obtain ML estimates by "extracting" the information contained in the observations when perceived as a sequence. The results of Oommen (November 2007) introduced the concepts of Sequence Based Estimation (SBE) for the Binomial distribution, where the authors derived the corresponding MLE results when the samples are taken two-at-a-time, and then extended these for the cases when they are processed three-at-a-time, four-at-a-time etc. This current paper generalizes these results for the multinomial case. The strategy we invoke involves a novel phenomenon called "Occlusion" that has not been reported in the field of estimation. The phenomenon can be described as follows: By occluding (hiding or concealing) certain observations, we map the estimation problem onto a lower-dimensional space, i.e., onto a binomial space. Once these occluded SBEs have been computed, we demonstrate how the overall Multinomial SBE (MSBE) can be obtained by mapping several lower-dimensional estimates, that are all bound by rigid probability constraints, onto the original higher-dimensional space. In each case, we formally prove and experimentally demonstrate the convergence of the corresponding estimates. The estimation methods proposed here have also been tested on real-life datasets from the UCI repository (Frank and Asuncion, 2013), and the accuracies obtained have been remarkable. We also discuss how various MSBEs can be fused to yield a superior MSBE, and present some potential applications of MSBEs. Our new estimates have great potential for practitioners, especially when the cardinality of the observation set is small.

1. Introduction

Since the sequence-based paradigm for supervised learning that is explored in this paper is relatively new, it is prudent that we first motivate it by explaining the perspective of this paper. Estimation methods generally fall into various categories, including the Maximum Likelihood Estimates (MLEs) and the Bayesian family of estimates (Bickel and Doksum, 2000; Casella and Berger, 2001; Duda et al., 2000; Fukunaga, 1990; van der Heijden et al., 2004) which are well-known for having good computational and statistical properties. Consider the strategy used for developing the MLE of the parameter of a distribution, $f_X(\theta)$, whose parameter to be estimated is θ. The input to the estimation process is the set of points/observations $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, whose elements are assumed to be generated independently and identically as per the distribution, $f_X(\theta)$. The process for computing the Maximum Likelihood (ML) estimate involves deriving the likelihood function, i.e., the likelihood of the distribution, $f_X(\theta)$, generating the sample points/observations $\mathcal{X}$ given θ, which is then maximized (by

☆The work of the first author was done while visiting at Myongji University, Yongin, Korea. The first author was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada and a grant from the National Research Foundation of Korea. This work was generously supported by the National Research Foundation of Korea funded by the Korean Government (NRF-2012R1A1A2041661).

* Corresponding author at: School of Computer Science, Carleton University, Ottawa, Canada K1S 5B6.
E-mail addresses: oommen@scs.carleton.ca (B.J. Oommen), kimsw@mju.ac.kr (S.-W. Kim).

http://dx.doi.org/10.1016/j.engappai.2017.05.001
0952-1976/

traditional optimization or calculus methods) to yield the estimate, $\widehat{\theta}_{MLE}$. The general characteristic sought for is that the estimate $\widehat{\theta}_{MLE}$ converges to the true (unknown) θ with probability one, or in a mean square sense. The Bayesian schemes work with a similar goal, except that rather than them using Likelihood functions, they compute the posterior distributions assuming that θ itself is a random variable with a known distributional form. Bayesian and ML estimates generally possess desirable convergence properties. Indeed, the theory of estimation has been studied for hundreds of years (Bickel and Doksum, 2000; Casella and Berger, 2001; Jones and Garthwaite, 2002; Ross, 2002; Shao, 2003; Sprinthall, 2002), and it has been the backbone for the learning (training) phase of statistical pattern recognition systems (Duda et al., 2000; Fukunaga, 1990; Herbrich, 2001; van der Heijden et al., 2004; Webb et al., 2002).

Traditionally, the ML and Bayesian estimation paradigms work within the model that the data, from which the parameters are to be estimated, is known, and that it is treated as a *set*. The position that we respectfully submit is that traditional ML and Bayesian methods ignore and discard[1] valuable *sequence*-based information. The goal of this paper is to "extract" and "utilize" the information contained in the observations when they are perceived *both as a set* and in *their sequence of appearance*. Put in a nutshell, this paper deals with the relatively new field of sequence-based estimation in which the goal is to estimate the parameters of a distribution by maximally "squeezing" out the *set*-based and *sequence*-based information latent in the observations.

The consequences of solving this problem are potentially many. Estimation, as researchers in almost all fields of science and engineering will agree, is a fundamental issue, in which the practitioner is given a set of observations involving the random variable, and his task is to estimate the parameters which govern the generation of these observations. Since, by definition, the problem involves random variables, decisions, predictions, regressions and classification related to the problem are, in some way, dependent on the practitioner obtaining reliable estimates of the parameters that characterize the underlying random variables. If we are able to obtain reliable estimates of the parameters under investigation by utilizing the set-based *and* sequence-based information, this could potentially have advantages in all of the above-mentioned fields.

More specifically, suppose that the user received $\mathcal{X}$ as a sequence of data points as in a typical real-life (or real-time) application such as those obtained in a data-mining domain involving sequences, or in data involving radio or television news items (De Santo et al., 2004). The question that we have investigated is the following: "Is there any information in the fact that in $\mathcal{X}$, x_i specifically precedes x_{i+1}?". Or in a more general case, "Is there any information in the fact that in $\mathcal{X}$, the **sequence** $x_i x_{i+1} \ldots x_{i+j}$ occurs $n_{i,i+1,\ldots i+j}$ times?". Our position, which we proved in Oommen (November 2007) for binomial random variables, is that even though $\mathcal{X}$ is generated by an i.i.d. process, there is information in these pieces of sequential data which can be "maximally" utilized to yield the so-called family of Sequence Based Estimators (SBEs). The problem was initially studied in Oommen (November 2007), but only for the case of binomial random variables.

In probability theory and statistics, for i.i.d. processes, the observation sample X has the probability given by the product $Pr(X) = Pr[x_1] \cdot Pr[x_2] \ldots Pr[x_N]$. This quantity is invariant with respect to arbitrary *permutation* of X. From that perspective, it can *appear* as if the formulation of the principle of SBEs is incorrect and misleading. But the fact is that the SBEs do not use the permutation-based sequential information; they, rather, only utilize the *co-occurrence* properties of different outcomes in the observation sample. Unlike the ML estimates of the probabilities $\{s_i\}$, which are simply given by the relative frequencies of $\frac{n_i}{N}$, the SBEs are computed alternatively from the relative frequencies of different non-overlapping subsequences of length two, three or longer.

Thus, for example, the probability $Pr[< a, b >]$, of a subsequence $<a, b>$, is given by the product $s_a s_b$. Clearly, the corresponding ML estimate is given by the relative frequency $\frac{n(a,b)}{N/2}$ of the occurrence of the subsequence $<a, b>$ in the set of all non-overlapping subsequences of length two.[2] From this viewpoint, the SBEs utilize the fact that the frequencies $n(a, b)$ are actually "constrained" by the given marginal frequencies, $\{n(i)\}$. Obviously, the method does not use the *overall* sequential information except the frequencies $n(a, b)$ within the window of interest.

As far as we know, apart from the results in Oommen (November 2007), there are no other reported results which utilize sequential information in obtaining such estimates. Unlike the results of Oommen (November 2007), though, we will deal with multinomial[3] and not binomial random variables. Also, as highlighted in Oommen (November 2007), unlike the use of sequence information in syntactic pattern recognition, grammatical inference and in modeling channels using Hidden Markov Models (which involve estimating the bigram and n-gram probabilities of *dependent* streams of data (Bunke, 1993; Duda et al., 2000; Friedman and Kandel, 1999), in our case, we assume that the elements in the stream of data, $\mathcal{X}$, occur *independently*, and yet have information not utilized by traditional MLE schemes.

1.1. Advantages of having multiple estimates

If the MLE and any SBE of the parameter θ converge to the *same true, unknown, value*, what then is the advantage of having multiple estimates? The answer lies simply in the fact that although the traditional MLE and the SBEs converge *asymptotically* to the same value, they all have *completely different* values. This is reminiscent of the fairy tale of the seven blind men who each described an elephant with completely different descriptions. While each of the descriptors is, in and of itself, not fully accurate, the composite picture can be rendered accurate. This is all the more true because the information used in procuring each of these estimates is, in one sense, "orthogonal". Further, since the convergence properties of MLEs is asymptotic, one can glean and effectively utilize other information when the number of samples examined is "small". Thus, one can use each of these estimates to get different classification boundaries. Ideally, though, one would see a fusion or combination of these estimates or their respective classifiers.

A concrete example involving a real-life example is given in Oommen (November 2007). Our work of Oommen (November 2007) also discussed various potential applications on SBEs.

1.2. Contributions of the paper

The contributions[4] of this paper can be catalogued as follows:

[1] This information is, of course, traditionally used when we want to consider *dependence* information, as in the case of Markov models and n-gram statistics.

[2] The computation of the ML estimates from overlapping subsequences requires examining a larger number of such overlapping pairs. Although this estimate is an approximation, the rationale for employing it is explained later.

[3] Multinomial distributions typically deal with *counting* the number of several alternative events in a number of independent trials. On the other hand, *Categorical* distributions are about the possible results of each of these events. In that sense, we are dealing with the categorical distribution inferred from the multinomial events. For the rest of the paper, we take the freedom to refer to these estimates as MSBEs.

[4] An abstract of a talk on MSBEs was included in the *Proceedings of the 2016*