A Faster Reliable Algorithm to Estimate the p-Value of the Multinomial llr Statistic

Uri Keich and Niranjan Nagarajan

Department of Computer Science, Cornell University, Ithaca, NY-14850, USA
{keich,niranjan}@cs.cornell.edu

Abstract. The subject of estimating the p-value of the log-likelihood ratio statistic for multinomial distribution has been studied extensively in the statistical literature. Nevertheless, bioinformatics laid new challenges before that research by often concentrating its interest on the "thin tail" of the distribution where classical statistical approximation typically fails. Hence, some of the more recent development in this area have come from the bioinformatics community ([5], [3]).
Since algorithms for computing the exact p-value have an exponential complexity, the only generally applicable algorithms for reliably estimating the p-value are lattice based. In particular, Hertz and Stormo have a dynamic programming algorithm whose complexity is $O(QKN^2)$, where Q is the size of the lattice, K is the size of the alphabet and N is the size of the sample. We present a new algorithm that is practically as reliable as Hertz and Stormo's and has a complexity of $O(QKN\log N)$. An interesting feature of our algorithm is that it can guarantee the quality of its estimated p-value.

1 Introduction

The subject of goodness-of-fit tests in general and of using the (generalized) log-likelihood ratio (*llr*) statistic, in particular, is of great importance in applications of statistics. In many applications, an important question to answer is how unlikely is it that an observed sample came from a particular multinomial distribution (H_0)? But in order to answer this question, we first need to quantify the similarity level between the observed sample distribution and the null distribution. The llr statistic, G^2 (defined below) is a popular measure as it is provably optimal under some conditions. Indeed, it is so popular that it has several other names which are more commonly used in the information theory and bioinformatics community: entropy distance, relative entropy, information content, Kullbak-Leibler divergence etc., all of which (upto a factor of N) stand for $I = G^2/2$, where

$$I = \sum_k X_k \log\left(X_k/(N\pi_k)\right) \quad [1],$$

for a null multinomial distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and a random sample $\boldsymbol{X} = (X_1, \ldots, X_K)$ of size $N = \sum_k X_k$. Note that $I = 0$ if and only if the empirical

[1] One can readily show that $G^2 = 2I$ is a generalized llr (e.g. [12]).

I. Jonassen and J. Kim (Eds.): WABI 2004, LNBI 3240, pp. 111–122, 2004.