Signal Processing 141 (2017) 322–330

Contents lists available at ScienceDirect

Signal Processing

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

Constructive minimax classification of discrete observations with arbitrary loss function

Lionel Fillatre

Université Côte d'Azur, CNRS, I3S, France

ARTICLE INFO

Article history:
Received 15 March 2017
Revised 9 June 2017
Accepted 14 June 2017
Available online 21 June 2017

Keywords:
Multiple hypothesis testing
Statistical classification
Minimax test
Linear programming

ABSTRACT

This paper develops a multihypothesis testing framework for calculating numerically the optimal minimax test with discrete observations and an arbitrary loss function. Discrete observations are common in data processing and make tractable the calculation of the minimax test. Each hypothesis is both associated to a parameter defining the distribution of the observations and to an action which describes the decision to take when the hypothesis is true. The loss function measures the gap between the parameters and the actions. The minimax test minimizes the maximum classification risk. It is the solution of a finite linear programming problem which gives the worst case classification risk and the worst case prior distribution. The minimax test equalizes the classification risks whose prior probabilities are strictly positive. The minimax framework is applied to vector channel decoding which consists in classifying some codewords transmitted on a binary asymmetric channel. The Hamming metric is used to measure the number of differences between the emitted codeword and the decoded one.

1. Introduction

The problem of classifying discrete distributions often appears in engineering applications, including pattern recognition with discrete-valued data [1,2], sensor network with quantized observations [3,4] image processing [5–7], and channel decoding [8,9] among others. The goal of this work is to decide between K hypotheses $\mathcal{H}_1$, ..., $\mathcal{H}_K$ where the Probability Mass Function (pmf) of the observed data $\boldsymbol{x}$ depends on the known value of a certain parameter given the hypothesis. A decision error is measured with an arbitrary loss function which depends both on the true hypothesis and the chosen one. We assume that the prior probabilities of the hypotheses are unknown. This is a classical assumption when the prior knowledge of observations is insufficient.

1.1. Minimax classification

Contrary to a purely Bayesian criterion which needs a complete statistical description of the problem [10], the minimax criterion is well adapted to classification problems where the probability of each hypothesis is unknown. This criterion consists in minimizing the largest probability to make a classification error. The optimal test consists in choosing the maximum of weighted likelihood functions. The weights are generally very difficult to calculate [11], even in some simple cases. Furthermore, the minimax test may satisfy the equalization property, i.e., the worst classification errors are all equal, which is quite interesting in practice.

There are two main trends in literature to design minimax test. The first trend consists in calculating analytically the minimax test. On the first hand, the minimax test is studied in a general setting [10,11]. Although there is a vast literature, it is still difficult to find an algorithm which calculates the minimax test for a specific situation. For instance, the famous book [10] does not describe any algorithm to compute a minimax test. On the second hand, the minimax test is often established for a specific issue [12–16] but the algorithm can not be easily extended to an other observation model.

The second trend consists in computing numerically the minimax test [17–30]. The paper [17] is certainly the first to use of programming techniques for testing two composite hypotheses based on discrete random variables. The LP approach was already implicit in [18] with real-valued observations. It is shown in [19] how to use the simplex method for calculating minimax decisions functions. Duality theory was first used in [20] and for the general case of minimax tests in [21]. The results are extended to the more general class of most stringent tests in [22]. The paper [23] introduced a framework where the theory of infinite LP is applicable. The survey [24] gives an overview of these pioneering approaches. All the above mentioned papers are focused on the classification of only two hypotheses.

E-mail address: lionel.fillatre@i3s.unice.fr

http://dx.doi.org/10.1016/j.sigpro.2017.06.020

In the case of several hypotheses, the work [26] is devoted to solving general minimax problems by iteration methods. To solve a decision problem with an arbitrary loss function, numerical solutions involving nonlinear optimization to obtain the least favorable distribution have been studied in [27,31,32]. Obtaining the least favorable distribution simplifies the problem but it does not provide necessarily an equalizer minimax test, even when it exists. Indeed, in case of discrete observations, the equalization of the classification risks needs the randomization of the test, which is not obtained when only the least favorable distribution is computed. The work [25] showed that the theory of infinite LP can be exploited for multiple hypotheses testing problems but they do not propose any constructive algorithm to solve the problem. The case of several hypotheses is closely related to the problem of minimax estimation [33], except that the parameter space is generally not finite but continuous and compact in case of minimax estimation.

1.2. Discrete observations and finite linear programming

This paper is in favor of a "discretize-then-optimize" approach, i.e., the case of discrete observations can be interpreted as the discretization, or quantization, of continuous real observations. Discrete observations often occur in signal processing applications where the quantization of continuous values is necessary [34]. Digital communications and image processing are some fields where quantization is crucially important [35,36] to limit the size of the storage or to describe a digital content with only a few features. Wireless sensor networks are characterized by limited resources, such as energy and communication bandwidth. One way to save energy is to limit the data transmitted in the network by using quantized data [3,37,38]. More generally, the approach studied in this paper can be easily applied to any signal processing applications where data quantization is of interest. The way the data are quantized is out of the scope of this work.

Discrete observations naturally involve finite Linear Programming (LP). In fact, as described in Section 2, the decision function is then a vector of reals which makes possible the construction of a finite LP problem to compute the minimax test and the worst prior distribution. After discretization, the standard LP problem [39] can be solved using techniques for large-scale LP [40], e.g., interior point methods, Dantzig-Wolfe decomposition, etc. On the contrary, continuous observations lead to infinite LP because the decision function generally belongs to an infinite dimensional space, as shown for instance in [22]. Infinite LP has the advantage to fit a general case but it is numerically difficult to solve as shown for example in [41,42]. The main way to numerically solve an infinite LP consists in discretizing the problem or to discretize the solution of the problem if it is known. Alternative approaches for solving infinite LP consist in approximating the initial problem by a sequence of LP problems with finite dimensional spaces [42]. The main drawback of this alternative "optimize-then-discretize" approach would be to develop ad-hoc optimization algorithms.

1.3. Contributions of the paper

The approach proposed in this paper is based on [22] where the author solves a LP problem to calculate the minimax test between only two hypotheses (binary classification). The paper [22] does not consider any loss function; only the probability of misclassification is studied. It deals with hypotheses which can be composite, i.e., each hypothesis may refer to an infinite number of statistical models. It is focused on continuous observations and it studies the minimax test as the solution of an infinite LP problem. It also proves a weak duality theorem between the primal infinite LP problem and its dual. The solution of the dual LP problem gives the worst case distribution of the minimax test. All the results proposed in [22] are theoretical and no algorithm is proposed, or can be easily derived, to compute the minimax test. The case of discrete observations is just very briefly introduced as a motivation of the general study. This paper extends [22] to the multiple hypothesis framework (K-ary classification with $K \geq 3$) and to an arbitrary loss function, i.e., it considers that the classification risk between a couple of hypotheses can change with respect to the involved couple of hypotheses. It only considers simple hypothesis: each hypothesis refers to only one statistical model. It is focused on discrete observations in order to make tractable the computation of the minimax test.

The first contribution of this paper is the design of a minimax classification test between multiple hypotheses as the solution of a finite LP problem, called the primal problem, when the observations are discrete and the loss function is arbitrary. This contribution is summarized in Theorem 2. The explicit calculation of the randomized minimax test makes it possible to equalize the classification risks, which is discussed in Corollary 1. This equalization of the classification risks is generally not fulfilled by a Bayes test because it depends on the worst case distribution.

The second contribution is the computation of the worst case distribution, also called the least favorable prior, which is obtained as the solution of the dual LP problem. The minimax test is then expressed as the maximum of weighted likelihood functions, i.e., it is a Bayesian test associated to the worst case weights. This contribution is summarized in Theorem 3. The calculation is very accurate since there is no need of a stopping criterion to halt the algorithm.

Finally, the minimax test is applied to noisy channel decoding. The Hamming metric is used to measure the number of differences between the emitted codeword and the decoded one. Assuming that the channel and the codebook are known but not the probabilities of each codeword, it is shown that the minimax test outperforms the conventional Maximum Likelihood (ML) decoder, also known as the Multiple Generalized Likelihood Ratio Test (MGLRT), which assumes an uniform prior over the codebook. The ML decoder is clearly suboptimal in case of the binary asymmetric channel when the prior distribution of the codewords is not uniform. It should be noted that the optimality of the minimax test is non-asymptotic and it is different from the random coding sense usually employed in channel decoding.

1.4. Organization of the paper

The paper is organized as follows. Section 2 describes the statistical framework, including the presentation of the minimax criterion and the LP problem whose solution is the minimax test. Section 3 studies the solution of the LP problems, both the primal and the dual ones, which lead to the minimax test closed-form expression and the worst case distribution of the hypotheses. Section 4 shows the relevance and efficiency of the proposed test for noisy channel decoding. Finally, Section 5 concludes this paper.

The following notations are used throughout the paper. The notation $X \sim p$ means that X follows the pmf p. The expectation of the function $f(X)$ with respect to the distribution of X is denoted $\mathbb{E}^X[f(X)]$. If $X \sim p_{\boldsymbol{\theta}}$ follows the distribution $p_{\boldsymbol{\theta}}$ parametrized by a vector $\boldsymbol{\theta}$, then the expectation is denoted $\mathbb{E}^X_{\boldsymbol{\theta}}[f(X)]$. Lower-case and upper-case letters are for scalar variables or random variables, bold lower-case letters for column vectors, bold upper-case letters for matrices and calligraphic upper-case letters or upper-case Greek letters for sets. Transposition, the transformation of columns into rows in a vector $\boldsymbol{x}$, resp. a matrix $\boldsymbol{A}$, is denoted by $\boldsymbol{x}^\top$, resp. $\boldsymbol{A}^\top$.