VARIABLE PRUNING IN BAYESIAN SEQUENTIAL STUDY DESIGN

Gergely HAJÓS
Advisors: Péter ANTAL, Tadeusz DOBROWIECKI

I. Introduction

In this paper we present a full bayesian method to take in account costs and utility of data used in induction. First we introduce the stopping problem, then the multi-armed bandit problem, then the Markov decision process. After the theoretical foundations we briefly discuss the two research areas connected to these problems: the adaptive study design, and the active learning. Finally we present the extension of this adaptive study design in a multivariate context for the problem of feature subset selection (FSS). The application area is the sequential study design of partial genome screening studies (PGAS).

The *optimal stopping problem* is defined by a sequence of independent random variables $X_1, X_2, \ldots, X_i$ with known distribution and a score function $f(\cdot)$ which gives a score to the observed sequence of random variables ($y_i = f(X_1, X_2, \ldots, X_i)$). In each step i a decision is made whether to stop and get score y_i or continue observing. The objective is to maximize expected reward. A special case of the optimal stopping problem is the *classical secretary problem*, which is defined as follows: there is a secreterial position to fill; there are n applicants; at each interview the applicant is accepted or rejected; the rejected applicants can not be accepted later; only one person can be accepted; the applicants can be ranked unambiguously; if the best applicant is chosen the reward is one otherwise zero.

If we can control the type of observations beside sample size, then we can formulate the *multi-armed bandit problem* (K-bandit problem). It is defined by K independent random variables $X_1, X_2, \ldots, X_K$ with unknown distribution. In each step i one has to select and sample one of the random variables x_i; the objective is to maximize the sum of samples $y = \sum_{i=0}^{M} x_i$ for a fixed M.

The multi-armed bandit problem is formally equivalent to the one state Markov decision process. The general *Markov decision process* is defined by: set of states S; set of actions A; conditional probabilities $Pr(s_{t+1} = s' | s_t = s, a_t = a)$ which in state s gives the probability of next state s' selecting action a; and utility function $U(s_{t+1} = s', s_t = s)$ which gives a reward for state change from s to s'. If a finite M step is assumed a possible objective is to define a $d(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ decision function to maximize $y = \sum_{i=0}^{M} Pr(s'_i | s_i, a_i) U(s'_i, s_i)$, where $s'_i = d(s_i, a_i)$.

The modern large-scale formalization of this line of research led to the concept of active learning and adaptive (sequential) study design. In this paper we follow the terminology of sequential study design (SSD).

The objective of the *sequential study design* is to retrieve statistical information from data collected sequentially, given a cost, utility and a budget constraint for data collection. A possible application is for instance to check the effectiveness of drug treatment or find association between genetic factors and diseases, etc. On the field of sequential study design generally the standard hypothesis testing approach is used [1]. The aim is to collect the minimum amount of data necessary to make a decision between null and alternative hypothesis, since one wants to minimize the costs of measurements performed in the study design. In every step of the sequential study design the tests of hypothesises are performed on a set of samples, if one of the hypothesises are accepted the study design is stopped and a decision is made. If a decision can not be made due to lack of enough information the study design continues and more data is collected.

Instead of applying the standard statistical approach in this paper we present a multivariate extension

based on Bayesian networks. In every step of the study design we first approximate the utility of computed results based on available data, second we predict future data based on available data using Bayesian model averaging over Bayesian networks. With the help of future data we predict the utility of the continuation of the study design. If the predicted utility of continuation is bigger than the utility based on available data then the sequential study design is continued, otherwise stopped and the last computed results are reported.

Beside selecting the sample size to minimize cost, in this paper we present an *active learning* approach to reduce the number of variables in every step of the sequential study design [2]. Since the algorithm in every step narrows down the set of investigated variables, in the subsequent step just the last (narrowest) set is used for further analysis. In this way in every subsequent step less and less measurements are performed.

In the case of partial genome association studies (PGAS) contrary to genome-wide association studies (GWAS) only partial information is available about the genome of the participants. In PGAS, we attempt to discover from subsequent measurements of well-selected blocks of variables the relevant genetic factors for a given target set with interim analyzis and meta-analyzis of the available aggregated data sets in order to interpret and guide further measurements (see Fig. 1). The phases are shown in Fig. 1, starting with the GWAS layer and the application of gene prioritization systems for the subjective, knowledge-rich initiation of our pruning process [3].

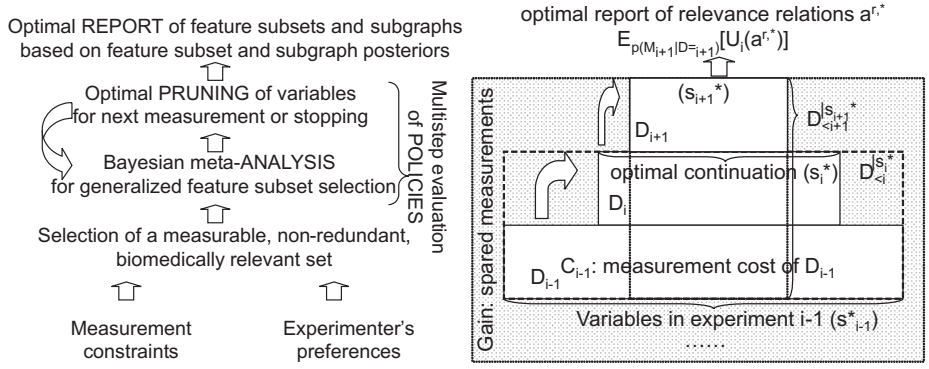

Figure 1: Left: The phases of sequential study design. Right: The main steps of Bayesian pruning in adaptive study design (for notation see Section II.)

In our approach we target the feature subset selection (FSS) problem in the Bayesian context. We apply complex Bayesian network based structural features in the analysis, specifically Markov Blanket Memberships (MBM), Markov Blanket Sets (MBS), and Markov Blanket Graphs (MBGs) [4]. The applied Bayesian multilevel relevance analysis means a multivariate approach to discover relevant sets of variables together with their interactions (conditional and contextual relevance), in correspondence we work with multivariate preferences (utilities). We evaluate typical policies in association studies based on interim Bayesian meta-analysis, and also the performance of a one-step look ahead approximation of the expected value of experiments in the full Bayesian approach. Our application domain is the investigation of genetic background of asthma using PGAS, where the sample collection and genotyping costs are considerable, and a multivariate approach is essential due to complex, weak interactions behind multifactorial diseases.

II. BAYESIAN SEQUENTIAL STUDY DESIGN AND VARIABLE PRUNING

Since in case of variable pruning beside the selection of the number of samples, we also narrow down the number of variables, we assume the following: a prior $p(M)$ for generative models;variable set S_i, where S_i contains only variables present in step i due to the reduction of variables; a corresponding

likelihood $p(D_i|M)$ for the ith step and data set D where D_i represents the data set narrowed down to variable set S_i with the samples N_i collected in step i; a set of actions, continuing sequential study design consisting of both the selection of S_{i+1}, N_{i+1} or reporting actions (stop experiment and report the last computed model); a context-free, e.g. timeless, cost $C_{N_i}^{S_i}$ of measuring (observing) D_i and a utility function $U(M_0, M^*)$ for stopping and reporting M^* in case of original model M_0. While $D_{<i}$ represents the data set narrowed down to variable set S_i with all the samples collected in steps $< i$ ($N_{<i} = \sum_{j=0}^{i} N_i$).

In the optimal Bayesian approach, at step $1 < i$ one possibility is to stop and report the optimal maximal utility model M^* with utility

$$U_i^{\text{report}} = E_{p(M|D_{<i})}[U(M, M^*)] - \sum_{j=1}^{i-1} C_{N_j}^{S_j}. \tag{1}$$

The other option is to continue by selecting the next, optimal experiment defined by selection of N_i, with utility

$$U_i^{\text{cont}} = U(N_i) - \sum_{j=1}^{i-1} C_{N_j}^{S_j}, \tag{2}$$

where $U(N_i)$ denotes the expected utility of experiment. In case of a decision problem with finite horizon, backward induction can be applied to calculate Eq. 2. The exponential number of potential future subsequent data makes however the estimation of these expectations computationally prohibitive (for nonmyopic evaluation of the value-of-information, see [5, 6]). The one-step, myopic approximation of $U(N_i)$ is as follows

$$U(N_i) \approx E_{p(D_i|D_{<i})}[E_{p(M|D_{\leq i})}[U(M, M^*)]], \tag{3}$$

which means that after the first step, the optimal Bayesian decision, (reporting M^* or continuing with measuring N_i) can be determined by comparing U_i^{report} to U_i^{cont}. For an overview and an upper bound for the expected value of an experiment, see [7]. Note that the framework of Markov decision processes is not directly applicable to this context, because of the dynamic state space.

III. MULTILEVEL ANALYSIS

We assume that there is a special set of target variables, and the goal is the identification of relevant variables, optimal sets of relevant variables, and their interactions (for an overview of FSS problem, see e.g. [8]). The goal of the analogous Bayesian FSS can be defined as the computation of the posteriors for the pairwise relations, relevant sets, and interactions, which can be formalized as the posteriors for MBM, MBS, MBG [4].

Respectively, we assume that the earlier utility function for the model can be decomposed into three parts, specifically for the MBM, MBS, and MBG levels. Note that given the utility function and the posterior over the feature space in step i, the expected utility of reporting a structural feature $\hat{f}$ can be computed and the feature value with maximal utility can be determined:

$$f^* = \arg\max_{\hat{f}} E_{p(f|D_{<i})}[U(\hat{f}|f)].$$

IV. RESULTS

One of the main motivations of the paper is to support more efficient measurements in partial genetic association studies. With this aim we evaluated three policies see Fig. 2. As the sensitivity curves

indicate after the first two pruning steps the external reference variables are still kept, e.g. in case of the greedy method the number of the pruned variables are 66 and 24, whereas all the MBS members are included. This means roughly that instead of measuring two times 116 variables, this allows the identification of the variables measuring only $116 + 66$ variables (i.e. saving one-quarter of the measurements).

For the artificial data set the policies could identify more than 69% of the relevant variables using < 152 measurements in three steps, which can be compared to a round robin scheme with $4 * 116$ measurements.

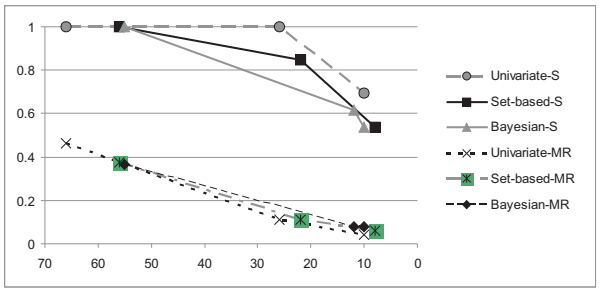

Figure 2: The horizontal axis shows the size of the selected sets in subsequent steps, the vertical axis shows the corresponding misclassification rate and sensitivity for univariate and set-based greedy, and the Bayesian policies in case of the artificial data set.

V. CONCLUSION

In the paper we investigated the decision support for sequential partial genetic association studies, which are essential methods to ensure high sample size for more targeted statistical analysis after GWAS-based explorations. We demonstrated that the multivariate generalization of the multi-armed bandit problem and budgeted learning — well-known in pharmacology and diagnostics — is viable in the sequential study design context as well, i.e. when both the predictive sampling and the evaluation are multivariate based on Bayesian networks. Preliminary results in an artificial context derived from real-world data indicate that significant saving is possible retaining high sensitivity.

References

[1] I. R. Konig and A. Ziegler, "Group sequential study designs in genetic-epidemiological case-control studies," *Hum. Hered.*, 56:63–72, 2003.

[2] J. Li, "Prioritize and select snps for association studies with multi-stage designs," *Journal of Computational Biology*, 15(3):241–257, 2008.

[3] S. Aerts, D. Lambrechts, S. Maity, P. V. Loo, B. Coessens, F. D. Smet, L. Tranchevent, B. D. Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature Biotechnology*, 24:537–544, 2006.

[4] P. Antal, A. Millinghoffer, G. Hullám, C. Szalai, and A. Falus, "A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction," *JMLR Proceeding*, 4:74–89, 2008.

[5] H. D., E. Horvitz, and B. Middleton, "An approximate non-myopic computation for value of information," in *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI'91)*, pp. 101–107. Morgan Kaufmann, 1991.

[6] W. Liao and Q. Ji, "Efficient non-myopic value-of-information computation for influence diagrams," *International journal of approximate reasoning*, 49:436–450, 2008.

[7] J. M. Bernardo, *Bayesian Theory*, Wiley & Sons, Chichester, 1995.

[8] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, 23(19):2507–2517, 2007.