The Budgeted Biomarker Discovery Problem: A Variant Of Association Studies

Sheehan Khan and **Russell Greiner**
Computing Science Department, University of Alberta
Edmonton, Alberta, Canada
{sheehank, rgreiner}@ualberta.ca

Abstract

In this paper we present the *budgeted biomarker discovery problem* as an alternative to the association studies traditionally used to identify biomarkers. We present several strong arguments to show why adopting this new problem will help solve issues in reproducibilty and understanding of association studies. Additionaly, we present several algorithms for this problem and show their performance on real metabolomic data and on synthetic data.

1 Introduction

Many researchers in bioinformatics are interested in discovering "differentially expressed" genes and proteins; also known as biomarkers. Researchers commonly use association studies to find these biomarkers. Unfortunately, the results of such association studies are often irreproducible, in that very different sets of genes appear as biomarkers in different studies done on the same phenotype (Ioannidis et al. 2009). Indeed, different labs will often produce different results, even using the same tissue samples (Yang et al. 2008).

While there are biological, technical and statistical explanations for this phenomenon (Ein-Dor, Zuk, and Domany 2006; Leek et al. 2010), part of the problem lies in the ambiguities inherent with the concept of association studies. Traditionally, association studies are viewed as multiple hypothesis testing problems, using some summary statistics to capture different trends in the data (Cui and Churchill 2003; Witten and Tibshirani 2007). However, the results of a study can vary wildly depending on which statistic is used (Boulesteix and Slawski 2009).

Another short coming of association studies is the issue of determining how different a feature must be for it to be considered significantly different – *i.e.*, when to reject the null hypothesis. Many researchers circumvent this issue by reporting the top k features (*e.g.*, k =50, or 100 or 1000). While such top lists may contain many features that are biologically interesting, no objective statements can be said of the quality of the list – *i.e.*, there is no biology-free validation for such lists. Other researchers prefer to control the false discover rate FDR (Reiner, Yekutieli, and Benjamini 2003; Storey and Tibshirani 2003). Note, however, that a list with 1 false positive in 10 purposed biomarkers is as good as one with 10 false positives in 100, as both have the same FDR $= 0.1$. However, as the goal of association studies is discovery, we should prefer the list with more actual biomarkers. Also note that perfect FDR is easy to achieve, by trivially declaring that *no* features are signifcantly different.

A final problem with association studies is that very few researchers release their data. Typically, they only provide basic summaries.[1] This may be a chicken-and-egg problem here, as researchers in the association studies community see no benefit to releasing their data, and thus people from the machine learning community are not motivated to develop methods to improve association studies due to lack of data sets for verifaction. We hope that our formalization will establish an initial foundation that both communities can extend.

Section 2 presents the *budgeted biomarker discovery problem*, and highlights why it is more precise, objective, and reproducible than standard association studes. Section 3 then describes a series of (increasingly effective) algorithms for solving this problem, with experimental comparisions (on both synthesized and real world data) in Section 4.

2 The Budgeted Biomarker Discovery Problem

In keeping with all the assumptions of association studies (Baldi and Long 2001; Efron et al. 2001; Smyth 2004), we first formalize things mathematically such that we may clearly define our *budgeted biomarker discovery problem*.

For each feature f_i (for $i = 1..N$), we let $\text{ex}(\, s,\, f_i\,) \in \Re$ be the (expression) value of feature f_i for subject s, whose label/phenotype is $\ell(\, s\,)$. We will asume only two classes, so $\ell(\, s\,) \in \{0, 1\}$.

For each f_i, we assume $\text{ex}(\, s,\, f_i\,) \;\sim\; \mathcal{N}\left(\mu_{i,\ell(\, s\,)}, \sigma_i^2\right)$ is normally distributed when conditioned on the class label (that is, the value of the binary phenotype); note the mean depends on that class, but the variance is common to both classes.

[1] One noteable exception is the microarray community, which posts their data sets to GEO – `http://www.ncbi.nlm.nih.gov/geo/`.

We assess each feature f_i based on its effect size $\Delta_i = \frac{\mu_{i,1}-\mu_{i,0}}{\sigma_i}$. If $\Delta_i = 0$ then we say the feature is *irrelevant*; and otherwise, $\Delta_i \neq 0$ means the feature is relevant – *i.e.*, is a biomarker. (We will later use '$+$' for biomarker and '$-$' for irrelevant.) We assume that the absolute values of the effect size $|\Delta_i|$ for biomarkers are random variables drawn from an exponential distribution with rate parameter λ. This assumption seems appropriate as the exponential distribution is the maximum entropy distribution if we assume we only know $E[|\Delta|]$ (Kapur 1989).

We consider an experimental setup where data is collected in a series of 'probe-pairs', which includes an observation of 1 specific feature from each of two patients: one in the '1' class and another in the '0' class. However, before we can collect any probes for feature f_i, we must first (pay to) collect C probes from samples of known concentrations to calibrate the machinery.

Definition [*Budgeted biomarker discovery problem*]

We are given (a) a set of features $\{f_i\}_{i=1}^N$, some of which are biomarkers '$+$' and others are not '$-$'; (b) a reward model $R_{TP}, R_{FN}, R_{FP}, R_{TN} \in \Re$, where

truth\prediction	$+$	0	$-$
$+$	R_{TP}	0	R_{FN}
$-$	R_{FP}	0	R_{TN}

and (c) a fixed budget of B on the total number of probes. An "assessment" is a function that maps each feature to $\{+, 0, -\}$ (where 0 means "undecided"). Our goal is to find a sequential probing strategy (spending at most B probes) and an assessment function to maximize our expected reward. Where appropriate, we may use an a priori probability π_i that feature f_i is a biomaker; if no information is available a priori we use $\pi_i = 0.5$ as the default. ■

This budgeted biomarker discovery problem has several advantages over the standard notion of an association study: (1) it has a clear and precise definition of biomarker versus irrelevant, as opposed to the subjective concept of "diffential expression"; (2) the fixed definition of biomarker versus irrelevant features should improve the reproducibility of the studies, as it removes the ambiguity on how "differential expression" is quantified; (3) it allows the discovery of both biomarker and irrelevant features, whereas association studies focused only finding biomarkers; (4) the reward model allows an "undecided" label, which means algorithms can avoid labeling features if they have low confidence in their assessments; and (5) it can use the known budget to produce algorithms that can avoid collecting data on features where the decision is obvious and hence spend more on features that have need more information, which can be more efficient than the the static experimental designs used in traditional association studies.

3 BBD Algorithms

Here we progressively build a series of BBD (budgeted biomarker discovery) algorithms, leading to the oSPRT (ordered sequential probability ratio test), which is our current best solution for the budgeted biomarker discovery problem.

Traditional: UNIFORM-BBD

The UNIFORM-BBD algorithm performs a traditional association study. It begins by selecting a subset of N' of the candidate features, and spending the probe budget uniformly accross them. Thus, each feature will be probed $n = \lfloor B/N' - C \rfloor$ times. Then it runs a standard two-sample t-test for each feature. If the resulting p-value is less than $p_{critical}$, the feature is labeled as '$+$', and otherwise it is labeled as '$-$'.

The results of this algorithm are very sensitive to the specific subset of features that are used. If it includes too many features, then n will be small, which makes it unlikely that any p-values are below $p_{critical}$. If it includes too few features, the possible score that the algorithm can achieve will be limited. We suggest using the priors $\{\pi_i\}$ and Monte-Carlo simulaiton to pick an appropriate set.

NAIVE-BBD

The NAIVE-BBD algorithm collects the same data as UNIFORM-BBD, but uses the basic modeling assumption that $|\Delta|$ is exponentially distributed for the '$+$' features to compute a posterior probability for each feature. NAIVE-BBD algorithm then returns the assessment to the features with highest expected evaluation. The posterior is computed by Eqn 3, which uses $f_t(x : \nu, \theta)$ to denote the non-central t-distribution with ν degrees of freedom and non-centrality parameter θ evaluated at x.

$$t = \left| \frac{\hat{\mu}_+ - \hat{\mu}_-}{\hat{\sigma}} \right|, \quad \nu = 2n - 2, \quad \theta = \Delta\sqrt{\frac{n}{2}}$$

$$f_-(t; n) = f_t(t : \nu, 0) + f_t(-t : \nu, 0) \quad (1)$$

$$f_+(t; n) = E_\Delta[f_t(t : \nu, \theta) + f_t(-t : \nu, \theta)] \quad (2)$$

$$p(+|t, n) = \frac{\pi f_+(t; n)}{\pi f_+(t; n) + (1-\pi) f_-(t; n)} \quad (3)$$

ORDERED-BBD

A problem with NAIVE-BBD is that the averaging in Eqn 2 places much of the distributional mass near the origin, which makes it harder to detect '$-$' features than '$+$' ones. A smarter algorithm can counteract this by noting that we expect $\hat{N}_- = \sum_{i=1}^{N'}(1-\pi_i)$ '$-$' features, and so we should expect, amongst those features, to see some smaller statistics than if we were observing just one feature.

This motivates the ORDERED-BBD algorithm, which agrees with NAIVE-BBD on the '$+$' features, but sorts the t-values, $t_{(1)} \leq \ldots t_{(\hat{N}_-)}$, so that ordered statistics can be used to detect the '$-$' features. For notational convenience, we drop the dependence on t and n, and use F_- to denote the CDF corresponding to the distribution in Eqn 1.

$$f_{(r)} = r\binom{\hat{N}_-}{r} \times (F_-)^{r-1} \times f_- \ldots \times (1-F_-)^{\hat{N}_- - r} \quad (4)$$

$$p(+|r) = \frac{\pi f_+}{\pi f_+ + (1-\pi) f_{(r)}} \quad (5)$$

ORDERED-BBD will label the feature with $t_{(r)}$ as '$-$' if Eqn 5 gives it sufficient confidence.

SPRT-BBD

Thus far, all the algorithms presented have worked with a static experimental design. We can potentially do better if we only probe a feature until we have sufficient confidence to label it as either '$+$' or '$-$'. It is straightforward to adopt the well studied sequential probability ratio test (SPRT) to this task (Wald 1945). The algorithm operates by picking the unassigned feature with the largest prior and collecting probes until a $\{+, -\}$ assessment can be made, then examines the feature with the next largest prior, and so forth.

OSPRT-BBD

Like the NAIVE-BBD algorithm, SPRT-BBD also suffers from the averaging effect in Eqn 2. This means SPRT-BBD needs more probes for the '$-$' features than the '$+$' ones. We can again use ordered statistics to remedy the situation, by tweaking SPRT-BBD to only collect probes until it has sufficent confidence to label the feature as '$+$' or it times out when n_{max} probes have been collected. Amongst the features that time out, the '$-$' ones will have i.i.d. t-statitics.

Once it exhausts the budget, OSPRT-BBD then applies the same ordered statistics techniques from Eqn 4 and Eqn 5 on the features that have timed out to label some of them as '$-$'.

We select an appropriate n_{max} value by running Monte-Carlo on synthetic data.

4 Experiments

We compare the algorithms on two datasets. The first is a real dataset that examines cancer cachexia from a metabolomic perspective (Eisner et al. 2011). The second data set is synthetic based on the modeling assumptions.

Real Data

The real data set contains expression values of 63 metabolites in 77 cancer patients, (47 with cachexia '1', 30 controls '0'). For the experiment, we do not know which of the features are actual biomarkers, so to set the ground truth we first compute p-values and effect sizes $\hat{\Delta}$ using all the data. We observe, that setting $p_{critical} = 0.05$ and $\lambda = 1$ both UNIFORM-BBD and NAIVE-BBD will produce the same assement for each of the 63 features. Thus, we set the $\{+, -\}$ labels for each feature based on those assessments. In the experiment, when an algorithm probes a feature, it will recieve a random observation from the available '1' patients, and a random observation from the available '0' patients. Lastly, we make the innocuous assumption that we have a constant prior $\pi = 0.3$ across all features to model the intuition that most features are not biomarkers.

To capture the budgeted nature of the problem, we set the budget to be $B = 1000$ probes and calibration cost $C = 10$ probes so that the algorithms can only collect half of the available data. We set the evaluation parameters to force a minimum confidence level of 80%: $R_{TP} = 1$, $R_{FP} = -4$, $R_{TN} = 1$, $R_{FN} = -4$.

To show the strength of OSPRT-BBD, we force it to use the best n_{max} found by Monte-Carlo in the following synthetic data experiment, while UNIFORM-BBD, NAIVE-BBD, and ORDERED-BBD are allowed to tune their parameters by training on the test data.

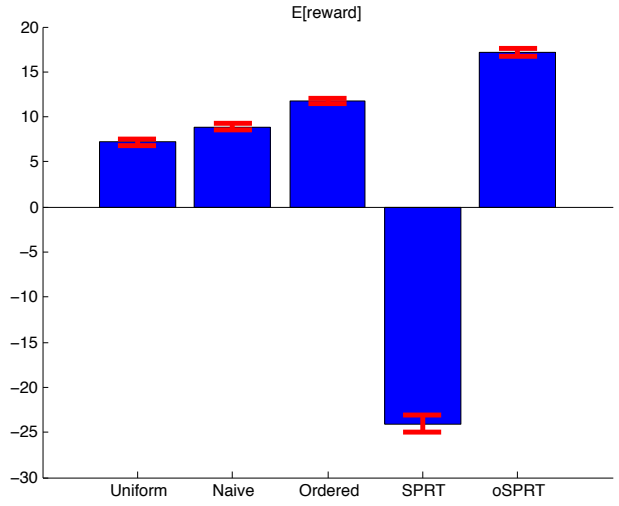

Figure 1: Plot of experimental results on the cancer cachexia data.

Figure 1 shows the results from this experiment. Firstly, we see UNIFORM-BBD performs the worst, because the assessments based on the p-values alone is the least robust to resampling noise. NAIVE-BBD and ORDRERED-BBD are more robust as their assessments directly consider the '$+$' vs. '$-$' nature of the problem. Note that ORDRED-BBD outperforms NAIVE-BBD based on its use of ordered statistics, and OSPRT-BBD does even better by also considering the budget as data is collected. The most interesting result is the total failure of SPRT-BBD. The reason for this failure is that, it is designed to make assessments as quickly as possible, and due to a mismatch between the model assumptions and the real data it over estimates its confidence. By quickly making slighlty bad decisions, SPRT-BBD saves budget to make even more bad decisions. Note, OSPRT-BBD does not have this problem as it must wait for n_{max} probes before it can potential label a feature as '$-$'.

Synthetic Data

We now repeat the real data experiment but use entirely synthetic data, with all the same parameters as previously assumed. To genereate the data for each feature f_i we first flip a coin with probability $\pi_i = 0.3$ to decide if it is a biomarker. If it is a biomarker, than we draw Δ_i from exponential distriobtion with $\lambda = 1$, if it is irrelevant than we set $\Delta_i = 0$. We then draw the probe data from the approriate normal distributions.

Figure 2 shows the results for the best parameter settings of each algorithm. We observe similar trends to the real data experiment but the scores are notably lower. Without access to more public datasets it is difficult to explain if this is a result of violation of the normal assumption, or an issue arising from re-sampling technique used to run the real data experi-

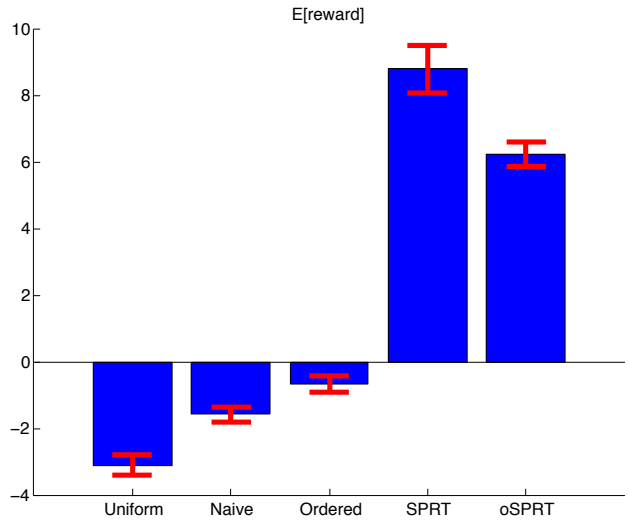

Figure 2: Plot of experimental results on synthetic data.

ment. The most important observation from this study is that on synthetic data the SPRT-BBD algroithm behaves exactly as it should and out performs the others. However, the advantage over OSPRT-BBD is small. Decomposing their results we find that OSPRT-BBD actually has higher true and negative positive rates, but achieves a lower score because it makes fewer '$+/-$' assessments.

5 Conclusions

This paper has motivated and presented the *budgeted biomarker discovery problem* (BBD) as a highly practical variant to standard association studies. Here, an algorithm can collect data in a series of probes, with the goal of assessing which features, in a pre-defined set of candidates, qualify as biomarkers. This BBD task extends association studies as it provides (1) a clear definition of what is and is not a biomarker; (2) an objective evaluation function for scoring the list of features labeled as biomarkers; and (3) a cost criterion for gathering the relevant information.

We presented 5 algorithms for tackling this BBD task and empirically compared them on both real and synthetic datasets. These empirical results confirm that it is advantagous to use an online decision process to quickly determine which features are obviously '$+$' or '$-$', which then means that more effort can be spent on the features that are hardest to assess. They also show that ordered statistics can be a powerful tool when simultaneosly assessing many features. We found that our OSPRT-BBD, which embodies both these principals, provides the best solution to this task.

References

Baldi, P., and Long, A. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17(6):509–19.

Boulesteix, A., and Slawski, M. 2009. Stability and aggregation of ranked gene lists. *Brief. Bioinform.* 10(5):556–68.

Cui, X., and Churchill, G. 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* 4(4):210.

Efron, B.; Tibshirani, R.; Storey, J.; and Tusher, V. 2001. Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc.* 96(456):1151–60.

Ein-Dor, L.; Zuk, O.; and Domany, E. 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.* 103(15):5923–8.

Eisner, R.; Stretch, C.; Eastman, T.; Xia, J.; Hau, D.; Damaraju, S.; Greiner, R.; Wishart, D.; and Baracos, V. 2011. Learning to predict cancer-associated skeletal muscle wasting from 1h-nmr profiles of urinary metabolites. *Metabolomics* 7(1):25–34.

Ioannidis, J.; Allison, D.; Ball, C.; Coulibaly, I.; Cui, X.; Culhane, A.; Falchi, M.; Furlanello, C.; Game, L.; Jurman, G.; Mangion, J.; Mehta, T.; Nitzberg, M.; Page, G.; Petretto, E.; and van Noort, V. 2009. Repeatability of published microarray gene expression analyses. *Nat. Genet.* 41(2):149–55.

Kapur, J. 1989. *Maximum-Entropy Models in Science and Engineering*. Wiley.

Leek, J.; Scharpf, R.; Bravo, H.; Simcha, D.; Langmead, B.; Johnson, W.; Geman, D.; Baggerly, K.; and Irizarry, R. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11(10):733–9.

Reiner, A.; Yekutieli, D.; and Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19(3):368–375.

Smyth, G. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3(1).

Storey, J., and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100(16):9440–5.

Wald, A. 1945. Sequential Tests of Statistical Hypotheses. *Ann. Math. Stat.* 16(2):117–186.

Witten, D., and Tibshirani, R. 2007. A comparison of fold-change and the t-statistic for microarray data analysis. Technical report.

Yang, H.; Harrington, C.; Vartanian, K.; Coldren, C.; Hall, R.; and Churchill, G. 2008. Randomization in laboratory procedure is key to obtaining reproducible microarray results. *PLoS One* 3(11).