Selective Model Averaging with Bayesian Rule Learning for Predictive Biomedicine

Jeya B. Balasubramanian, MS[1,2], Shyam Visweswaran, MD, PhD[1,2,3], Gregory F. Cooper, MD, PhD[1,2,3], Vanathi Gopalakrishnan, PhD[1,2,3]

[1]Department of Biomedical Informatics, [2]Intelligent Systems Program, [3]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA

Abstract

Accurate disease classification and biomarker discovery remain challenging tasks in biomedicine. In this paper, we develop and test a practical approach to combining evidence from multiple models when making predictions using selective Bayesian model averaging of probabilistic rules. This method is implemented within a Bayesian Rule Learning system and compared to model selection when applied to twelve biomedical datasets using the area under the ROC curve measure of performance. Cross-validation results indicate that selective Bayesian model averaging statistically significantly outperforms model selection on average in these experiments, suggesting that combining predictions from multiple models may lead to more accurate quantification of classifier uncertainty. This approach would directly impact the generation of robust predictions on unseen test data, while also increasing knowledge for biomarker discovery and mechanisms that underlie disease.

Introduction

Models that predict phenotypes and disease states from high-dimensional '-omic' datasets can lead to discovery of useful and predictive biomarkers. The typical approach for learning predictive models is to perform model selection wherein a single model is selected that summarizes the data well. However, when using real datasets there may be substantial uncertainty in choosing one model over all others, especially when the selected model is one of several models that all summarize the data more or less equally well. A sound approach in this situation is *Bayesian model averaging* (BMA) wherein the prediction for a test instance is obtained from a weighted average of the predictions of all possible models within a model space, with more probable models influencing the prediction more than less probable ones (Hoeting et al., 1999). Often in real datasets, the number of possible models is enormous, and averaging the predictions over all of them is infeasible. A practical approach is to average over a few good models, termed *selective BMA*, which serves to approximate the predictions that would be obtained from averaging over all models. The method that we describe in this paper performs selective BMA over a set of probabilistic rules as an approximation to complete BMA over all such rules.

In this paper, we extend a novel rule generation method called Bayesian Rule Learning (BRL), which identifies a single set of probabilistic classification rules learned from a training data set that can be applied to predict the class value on unseen test data. We perform selective BMA over the rule sets of BRL in order to account for model uncertainty. We compare this selective BMA approach to a model selection approach and report experimental results obtained from a range of biomedical datasets.

Background

In this section, we provide details of constrained Bayesian networks, the Bayesian scoring of models, the implementation of model selection in BRL, and the selective model averaging version of BRL (SMA-BRL).

Bayesian networks: A Bayesian network (BN) is a probabilistic graphical model that combines a graphical representation of the probabilistic dependencies between variables and the probabilistic parameters of the BN. The graphical structure is a directed acyclic graph (DAG), where the nodes represent the predictive variables and edges represent a (conditional) probabilistic dependency between corresponding variables. Absence of an edge indicates (conditional) probabilistic independence between the corresponding variables. The probabilistic parameters represent joint probability distributions over a set of predictive variables. BRL uses a Bayesian score (described below) to evaluate constrained BN structures (see Figure 1a). A complete decision tree (see Figure 1b) represents the parameters of the target node.

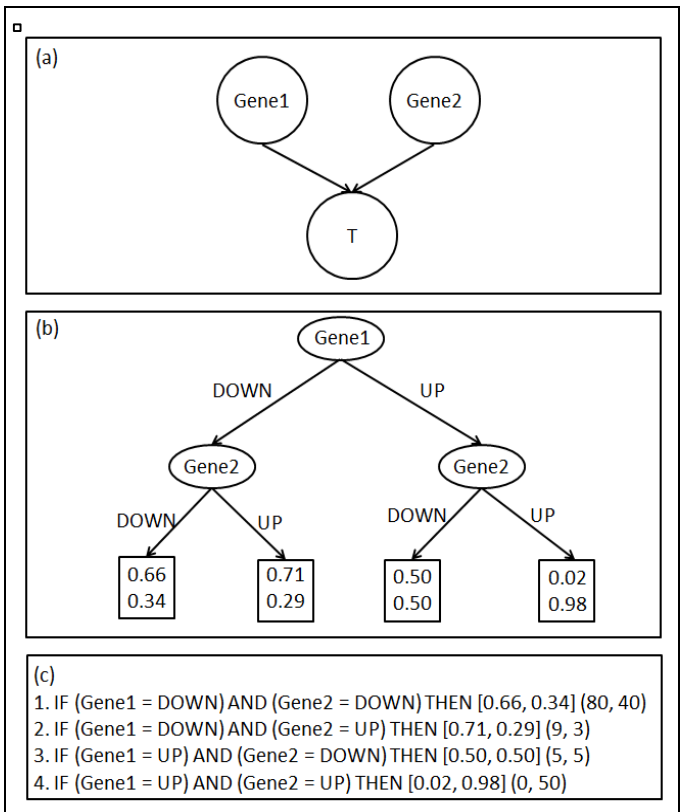

Figure 1: An example of a BN structure learned from BRL. Panel (a) displays a constrained BN structure (*S*) with two predictive variables, 'Gene1' and 'Gene2', as parents of the target variable *T*. The two predictive variables are binary with values of 'UP' and 'DOWN'. The target variable is binary having values 'Case' and 'Control'. Panel (b) shows the parameters for the target node as a complete decision tree. The interior nodes of the tree are the predictive variables (represented by ellipses) and the leaf nodes (represented by rectangles) show the probability distribution over *T*. Panel (c) shows the rule set inferred from the decision tree by BRL. Each rule antecedent is a path from a leaf to the root node. The consequent is the probability distribution of *T*. The following parentheses show the number of 'Case' instances and the number of 'Control' instances that match the antecedent, respectively.

This tree contains internal nodes that represent predictive variables and terminal nodes (leaves) that store the probability distribution over the target variable. Each leaf has a unique path to the root node. Each path represents a unique configuration of the parental states. Together, the leaves represent every possible parental state. BRL infers a set of rules (Figure 1c) from the decision tree. The rule set is the classifier model that describes the learned graphical structure and the probabilistic parameters. These rules are used to predict the target value for an unseen instance.

Bayesian score: BRL (Gopalakrishnan et al., 2010) learns a constrained BN structure (where a subset of the predictor variables have edges to the target) and evaluates it using a Bayesian score, which is proportional to the likelihood of the BN structure given the data. In this paper, we use the BDeu score (Heckerman, et al., 1995) to evaluate the BN structures. Equation 1 gives the BDeu score for the target node in the BN structure. Here, the symbol Γ represents the gamma function; j iterates through each of the q joint parental states of the target node in the BN rule-structure S; k iterates through each of the r states of the target node. N_{jk} is the number of instances (samples) in the dataset D in which the target has state k and parents of the target have state j. Here, $N_j = \sum_{k=1}^{r} N_{jk}$. The term α_0 is a user-defined parameter, which is called the *prior equivalent sample size* (*pess*). In this paper we set $\alpha_0 = 1$.

$$P(D|S) = \prod_{j=1}^{q}\left(\frac{\Gamma\left(\frac{\alpha_0}{q}\right)}{\Gamma\left(N_j+\frac{\alpha_0}{q}\right)}\cdot\prod_{k=1}^{r}\frac{\Gamma\left(N_{jk}+\frac{\alpha_0}{q\cdot r}\right)}{\Gamma\left(\frac{\alpha_0}{q\cdot r}\right)}\right). \qquad (1)$$

Methods

Algorithms: We re-implemented the beam search in the BRL, as shown in Figure 2. The beam is a priority queue of size *W*, which holds a set of BN structures ordered by the Bayesian score. The new implementation removes certain constraints imposed by the previous implementation, such that, we now search a larger space of models and we ensure that the final beam (of size *W*) returns each of the best *W* structures (according to the Bayesian score) evaluated by the search procedure.

Model selection in BRL returns a single model, *S*, from a total of *W* models that are generated from the training data, *D*, by the use of beam search. For a given vector of predictor variable values *X*, model *S* generates the posterior distribution of the target values $P(T \mid X, S)$. This does not account for the uncertainty of model *S* which is described by its posterior probability $P(S \mid D)$. In model selection, the selected model is assumed to have a posterior probability of 1. In reality, we are not certain that the model with the highest Bayesian score is indeed the data-generating model. Ideally, we should account for this uncertainty. In Bayesian model averaging, the predicted posterior distribution of the target is weighted by the uncertainty of the model, for all models in the model space. These terms are then summed to obtain the model averaged posterior distribution of the target. The cardinality of the

model space in BRL is $\sum_{b=0}^{B} \binom{n}{b}$, where n is the number of predictor variables and B is the maximum number of parents the target node can have. The total number of models grows rapidly in n. It is generally not feasible to enumerate every possible model. Instead, we make use of the W models already available from the existing beam search in the BRL. For model averaging, we average over these W models. This is called selective Bayesian model averaging. Equation 2 gives the average of the posterior distributions of the target node T, averaged over W models.

$$P(T|X) = \sum_{i=1}^{W} P(T|X, S_i) \cdot P(S_i|D) \quad (2)$$

The posterior probability of each model in Equation 2 is derived using Equation 3.

$$P(S_i|D) = \frac{P(D|S_i) \cdot P(S_i)}{\sum_{j=1}^{W} P(D|S_j) \cdot P(S_j)} \quad (3)$$

INPUT: A training dataset *D* with *m* instances, a set of *n* discrete predictor variables **X** = {X_1, X_2… X_n}, and a discrete target variable *T*. The maximum number of parents that the target node *T* can have is *max_parents*. The maximum number of rule structures that the beam holds is *W*.

OUTPUT: Returns the posterior distribution of the target node given an input vector of *n* discrete variables.

DEFINITIONS:
T is the target node;
Score(*S*) = *P*(*D* | *S*). This is the Bayesian score for rule structure *S* generated from dataset *D* (see Equation 1);
Q = Priority queue defined by the set {S_1, S_2… S_W} where *Score*(S_i) > *Score*(S_j), when $i < j$;
F = Priority queue of models;
V = Set of all variables in dataset *D*, except target variable *T*;
$\pi(S)$ denotes the parents in rule structure *S*;
max_parents = 8 (default);
W = 1000 (default);

ALGORITHM:
BRL_BeamSearch:

```
1.      Create structure S containing just target node T and place S onto Q
2.      WHILE (Q is not empty) DO:
3.              S_curr ← remove(Q)
4.              IF(F does not contain S_curr):
                        place S_curr onto F
                END-IF
4.              V' = V - π(S_curr) //V_i is a set of all variables not in S_curr.
5.              IF ( (V' ≠ ϕ) AND (|π(S_curr)| < max_parents ) THEN:
6.                      FOR-EACH (v in V') DO:
                                S_new ← Add v as a parent of T in S_curr.
                                IF(Q does not contain S_new):
                                        place S_new onto Q
                                END-IF
                                IF(F does not contain S_new):
                                        place S_new onto F
                                END-IF
                        END-FOR-EACH //Ends all specialization
8.                      Trim Q to the first W elements
                        Trim F to the first W elements
                END-IF
        END-WHILE       //End of beam search
9.      Return F
```

(a) BRL_ModelSelection:

```
1.      F = BRL_BeamSearch:
2.      S_best ← remove(F)
3.      S_best is used to predict T.
```

(b) BRL_SelectiveModelAveraging:

```
1.      F = BRL_BeamSearch:
2.      The W models in F are used to predict T (see Equation 2).
```

Figure 2: Algorithm for model selection and model averaging in BRL.

Biomedical datasets: We analyzed the performance of SMA-BRL and BRL on 12 publicly available biomedical datasets that are listed in Table 1. It has been shown that irrelevant variables tend to introduce noise during the model search process when there are high-dimensional biomedical data with a large number of predictor variables, but relatively few samples (Liu, et al. 2012). As an initial step, we therefore applied the Partitioning-based Adaptive Irrelevant Feature Eliminator (PAIFE) to remove irrelevant features. PAIFE deems a variable as 'unconditionally relevant' by using a univariate analysis that adaptively employs the chi-square test or the Fisher's exact test. PAIFE also detects 'conditionally relevant' variables from subsets of variables, where the relationship of the variable to the target variable is conditional over other variables. The variables that are neither conditionally nor unconditionally relevant were considered irrelevant and were removed from the dataset.

Table 1. The 12 biomedical datasets used for analysis. The first eleven are genomic and the twelfth one is proteomic. The data are identified with the 'Dataset ID'. The column 'P/D' describes the type of data as Prognostic (P) or Diagnostic (D). The '# V' column is the number of predictor variables originally in the dataset. The '$\#V_{PAIFE}$' column shows the number of variables selected by PAIFE. The 'Sample Class Distribution' shows the number of samples in each class in the dataset. The 'Reference' points to the relevant literature for the dataset.

Dataset	P/D	#V	$\#V_{PAIFE}$	Sample class distribution	Reference
1	D	6584	1972	40:21:00	(Alon, et al., 1999)
2	D	12582	2371	28:24:20	(Armstrong, et al., 2002)
3	P	5372	858	69:17:00	(Beer, et al., 2002)
4	D	7129	2288	47:25:00	(Golub, et al., 1999)
5	D	7464	1880	18:18	(Hedenfalk, et al., 2001)
6	P	7129	699	40:20:00	(Iizuka, et al., 2003)
7	D	2308	832	29:25:17:12	(Khan, et al., 2001)
8	D	7129	1927	58:19:00	(Shipp, et al., 2002)
9	D	10510	6713	52:50:00	(Singh, et al., 2002)
10	P	24481	4251	44:34:00	(Veer, et al., 2002)
11	D	7039	1230	35:04:00	(Welsch, et al., 2001)
12	D	70	15	139:66	(Bigbee, et al., 2012)

Experimental methods: We wanted to evaluate and compare the predictive performance of BRL and SMA-BRL, to quantify the change in predictive performance due to model averaging. We evaluated the two algorithms, over the 12 publicly available datasets, using 10 runs of 10-fold stratified cross-validation. For a given run, the mean performance (see below) over the 10 folds was derived. We used the average of those means as an estimate of the predictive performance of the algorithm for a given dataset.

Discretization: BRL and consequently SMA-BRL require discrete values for all the variables in the input dataset. The datasets that we analyzed (see Table 1) have continuous valued predictor variables, and a discrete target variable. In the 10 runs of 10-fold cross-validation, each fold was discretized using the efficient Bayesian discretization (EBD) method (Lustgarten, et al., 2011). EBD takes a parameter λ, which determines the expected number of cut-points for each variable. For our analysis, we set $\lambda = 0.5$.

Performance measure: The performance of the algorithms was evaluated using the percentage of the area under the ROC curve (AUC). The area under the ROC curve is typically used as a summary statistic of discrimination. The AUC is equivalent to the probability that a randomly chosen case from the negative class will have a smaller predicted probability of belonging to the positive class than a randomly chosen case from the positive class.

The average AUCs obtained from the two algorithms for each of the 12 datasets, over 10 runs of 10-fold stratified cross-validation, is analyzed using two statistical tests. We used the tests to check whether the difference between the performances of the two classifiers over the 12 datasets is non-random. The tests included (1) significance testing with the Wilcoxon paired-samples two-sided signed ranks test, and (2) effect size testing with paired-samples two-tailed t-test. We used the Statistics Toolbox from MATLAB to perform these tests (MATLAB and Statistics Toolbox Release 2013b, The MathWorks, Inc., Natick, Massachusetts, United States).

Results and discussion

The average AUCs obtained from the two algorithms for each of the 12 datasets is shown in Table 2. The result from the significance tests show that SMA-BRL is statistically significantly better than BRL based on these AUC values.

Table 2. Average AUCs obtained from BRL and SMA-BRL using 10 runs of 10-fold cross-validation for the 12 datasets described in Table 1. For each dataset, the result of the better performing algorithm is shown in bold. The last row shows the average from the 12 datasets and the standard error of mean (SEM).

Dataset	BRL	SMA-BRL
1	**99.50**	**99.50**
2	95.12	**95.67**
3	60.14	**60.25**
4	91.88	**93.82**
5	94.13	**100.00**
6	57.19	**58.13**
7	84.67	**86.55**
8	81.58	**82.87**
9	90.87	**90.95**
10	86.12	**86.50**
11	95.42	**97.92**
12	80.96	**82.28**
Average ± SEM	84.80 ± 3.90	86.20 ± 4.05

The non-parametric Wilcoxon paired-samples signed ranks test, with significance level $\alpha = 0.05$, shows that SMA-BRL performs statistically significantly better than BRL with a p-value of 9.765×10^{-4}. The paired-samples two-tailed t-test, with significance level $\alpha = 0.05$, also shows that SMA-BRL performs statistically significantly better than BRL with a p-value of 0.0122. The 95% confidence interval of the mean of the difference between the column values of BRL and SMA-BRL in Table 2, based upon the t-distribution is [-2.438, -0.372].

We observe that the difference between the average AUC, for BRL and SMA-BRL, across the 12 datasets is small. We also observe that for each of the 12 datasets we analyzed in this paper, SMA-BRL either obtains an equivalent or better average AUC performance than BRL. Note that the SMA-BRL uses the same search engine as the BRL. The BRL generates *W* models but only one is selected and used for inference on a test case. SMA-BRL makes use of all the *W* models for its inference. Therefore, SMA-BRL only requires an additional constant time operation during the model inference step. As a result, SMA-BRL returns a robust classifier, which is worth exploring, at very little additional computational cost. A limitation of SMA-BRL when compared to BRL is that the predictions based on SMA-BRL involve the weighted inference of *W* probabilistic rules, which is more complex to understand than the inference of a single rule in BRL.

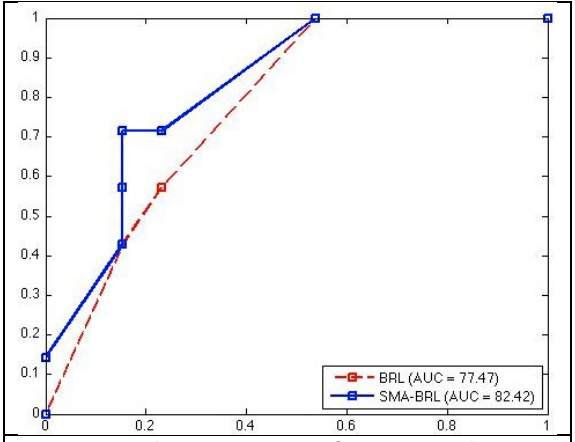

Figure 3: The ROC curve for BRL and SMA-BRL, on one training fold of dataset 12 (Table 1).

Case Study: We examined the models learned by BRL and SMA-BRL on one of the training folds of dataset 12 (see Table 1). The ROC curve of the models is shown in Figure 3. The AUC of the BRL model is 77.47 and of the SMA-BRL model is 82.42. The BRL model included three biomarkers (MIF, Thrombos, and SAA) and the SMA-BRL model, in addition to the three biomarkers, included five more biomarkers (IL-8, IGFBP-1, PROLACTI, TTR, and RANTES). In future work, we plan to study the relative importance of these variables and their biological significance.

Conclusion

SMA-BRL accounts for model uncertainty, which the model selection method BRL ignores. SMA-BRL generates robust predictions from a committee of plausible models. The strong theory underlying Bayesian model averaging is supported by the results from our analysis of 12 datasets. Moreover, since SMA-BRL only averaged over the models encountered in the BRL search, the computational time complexity of the two algorithms is almost identical. Thus, the improved results achieved with SMA-BRL are obtained essentially for free. Overall, these results support using model averaging when predicting outcomes in biomedical datasets that are similar to the 12 datasets analyzed in this paper.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments that helped tailor the paper for the intended audience. The authors gratefully acknowledge grant number R01-LM010950 from the National Library of Medicine. VG was funded in part by grants R01GM100387 and P50CA090440 from the National Institutes of Health. GFC was funded in part by NIH grant R01LM010020 and NSF grant IIS0911032. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382-401.
2. Gopalakrishnan, V., Lustgarten, J. L., Visweswaran, S., & Cooper, G. F. (2010). Bayesian rule learning for biomedical data mining. *Bioinformatics*, 26(5), 668-675.
3. Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning, 20(3), 197-243.
4. Lustgarten, J. L., Visweswaran, S., Gopalakrishnan, V., & Cooper, G. F. (2011). Application of an efficient Bayesian discretization method to biomedical data. *BMC Bioinformatics*, 12(1), 309.
5. Liu, G., Kong, L., & Gopalakrishnan, V. (2012). A Partitioning Based Adaptive Method for Robust Removal of Irrelevant Features from High-dimensional Biomedical Datasets. *Proceedings of the AMIA Summits on Translational Science*, 2012, 52.
6. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12), 6745-6750.
7. Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., Den Boer, M. L., Minden, M. D., … Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30, 41–47.
8. Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., … Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8, 816–824.
9. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.
10. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., … Sauter, G. (2001). Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, 344, 1–6.
11. Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., … Hamamoto, Y. (2003). Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, 361, 923–929.
12. Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., … Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*, 673–679.
13. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8, 68–74.
14. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., … Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203–209.
15. Veer, L. van't, Dai, H., & Vijver, M. Van De. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*.
16. Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., ... & Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences*, 98(3), 1176-1181.
17. Bigbee, W. L., Gopalakrishnan, V., Weissfeld, J. L., Wilson, D. O., Dacic, S., Lokshin, A. E., & Siegfried, J. M. (2012). A multiplexed serum biomarker immunoassay panel discriminates clinical lung cancer patients from high-risk individuals found to be cancer-free by CT screening. *Journal of Thoracic Oncology*, 7(4), 698.