A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets

Xia Jiang,[1] Binghuang Cai,[1] Diyang Xue,[1] Xinghua Lu,[1] Gregory F Cooper,[1] Richard E Neapolitan[2]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-002358)

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

Correspondence to
Dr Xia Jiang,
Department of Biomedical Informatics, University of Pittsburgh, Room 518, 5607 Baum Boulevard, Pittsburgh, PA 15206, USA;
xij6@pitt.edu

Received 14 September 2013
Revised 20 February 2014
Accepted 14 March 2014
Published Online First 15 April 2014

ABSTRACT

Objective The objective of this investigation is to evaluate binary prediction methods for predicting disease status using high-dimensional genomic data. The central hypothesis is that the Bayesian network (BN)-based method called efficient Bayesian multivariate classifier (EBMC) will do well at this task because EBMC builds on BN-based methods that have performed well at learning epistatic interactions.

Method We evaluate how well eight methods perform binary prediction using high-dimensional discrete genomic datasets containing epistatic interactions. The methods are as follows: naive Bayes (NB), model averaging NB (MANB), feature selection NB (FSNB), EBMC, logistic regression (LR), support vector machines (SVM), Lasso, and extreme learning machines (ELM). We use a hundred 1000-single nucleotide polymorphism (SNP) simulated datasets, ten 10 000-SNP datasets, six semi-synthetic sets, and two real genome-wide association studies (GWAS) datasets in our evaluation.

Results In fivefold cross-validation studies, the SVM performed best on the 1000-SNP dataset, while the BN-based methods performed best on the other datasets, with EBMC exhibiting the best overall performance. In-sample testing indicates that LR, SVM, Lasso, ELM, and NB tend to overfit the data.

Discussion EBMC performed better than NB when there are several strong predictors, whereas NB performed better when there are many weak predictors. Furthermore, for all BN-based methods, prediction capability did not degrade as the dimension increased.

Conclusions Our results support the hypothesis that EBMC performs well at binary outcome prediction using high-dimensional discrete datasets containing epistatic-like interactions. Future research using more GWAS datasets is needed to further investigate the potential of EBMC.

BACKGROUND

The advances in high-throughput technologies have provided us with abundant data resources. For example, we have accumulated a huge number of single nucleotide polymorphism (SNP) datasets as a result of genome-wide association studies (GWAS). An SNP results when a nucleotide that is typically present at a specific location on the genomic sequence is replaced by another nucleotide.[1] These high-dimensional GWAS datasets can list over a million SNPs, and whole genome sequencing can produce datasets with hundreds of millions of SNPs.[2]

By looking at single-locus associations using these datasets, researchers have identified over 150 risk loci associated with 60 common diseases and traits.[3–6] However, it is likely that the discovery of loci with significant main effects reveals only a small fraction of the undiscovered genetic risk of many common diseases.[7–11] That is, much of the genetic risk might be due to undiscovered epistatic interactions, which are interactions by which several genes combined affect disease, and the net effect on phenotype cannot be predicted by simply combining the effects of the individual loci.

In light of its importance, researchers have sought to detect epistasis using genomic data. As standard techniques such as linear regression may not work well because the interactions are non-linear, other techniques were explored. One well-known technique is multifactor dimensionality reduction (MDR).[12] MDR combines two or more variables into a single variable and has been successfully applied to detect epistatic interactions in hypertension,[13] sporadic breast cancer,[14] and type II diabetes.[15] Jiang *et al*[16] evaluated the performance of 22 Bayesian network (BN) scoring criteria and MDR when scoring candidate interactions. They found that several of the BN scoring criteria performed substantially better than other scores and MDR.

A second difficulty when learning epistasis from high-dimensional datasets concerns the curse of dimensionality. That is, there are too many SNPs to look at all possible interactions. Therefore, researchers worked on developing heuristic search methods. Traditional techniques such as logistic regression (LR),[17] LR with an interaction term,[18] penalized LR,[19] and Lasso[20 21] were applied to the task. Other techniques included full interaction modeling,[22] using information gain,[23 24] SNP Harvester,[25] permutation testing,[26 27] the use of ReliefF,[28 29] random forests,[30] predictive rule inference,[31] a variational Bayes algorithm,[32] Bayesian epistasis association mapping,[33 34] maximum entropy conditional probability modeling,[35] a Markov blanket method,[36] an ensemble-based method that uses boosting,[37] and greedy search using BN scoring criteria.[38]

Our goal is not only to discover interactions from high-dimensional datasets, but also to use the knowledge learned to perform classification and prediction in a clinical setting. Many binary prediction methods are available, but currently researchers have little reason for choosing one over the other. Just as standard techniques may not work well for

To cite: Jiang X, Cai B, Xue D, *et al*. *J Am Med Inform Assoc* 2014;**21**: e312–e319.

learning epistatic interactions from high-dimensional genomic datasets, standard prediction methods may not do well at prediction using these datasets. Several BN-based prediction algorithms have been developed. The most basic is the naive Bayes (NB) classifier.[39] Variations of it that were designed to handle high-dimensional data include model averaging NB (MANB),[40 41] feature selection NB (FSNB),[40] and the efficient Bayesian multivariate classifier (EBMC).[42]

In an effort at consolidating these efforts and determining which methods work well for this problem, we compare the prediction performance of the BN-based methods to LR, support vector machines (SVM),[43] Lasso,[20] and extreme learning machines (ELM).[44]

OBJECTIVE

The objective of this investigation is to evaluate the performance of binary prediction methods when predicting disease status using high-dimensional discrete genomic data. Our *central hypothesis* is that the BN-based method EBMC will do well at this task because EBMC builds on BN-based methods that have performed well at learning epistatic interactions.

METHOD

We first review the methods we applied.

Bayesian networks

BNs[39 45–48] are used for uncertain reasoning and machine learning in many domains, including biomedical informatics.[49–54] A BN consists of a directed acyclic graph (DAG) $G = (V, E)$, whose nodeset V contains random variables and whose edges E represent relationships among the random variables. A BN also includes a conditional probability distribution of each node $X \in V$ given each combination of values of its parents. Each node V in a BN is conditionally independent of all its non-descendents given its parents in the BN. Figure 1 shows a causal BN modeling the situation that SNP rs7115850 and the APOE gene both have a causal inference on late onset Alzheimer's disease (LOAD) and LOAD has a causal influence on mental dexterity.

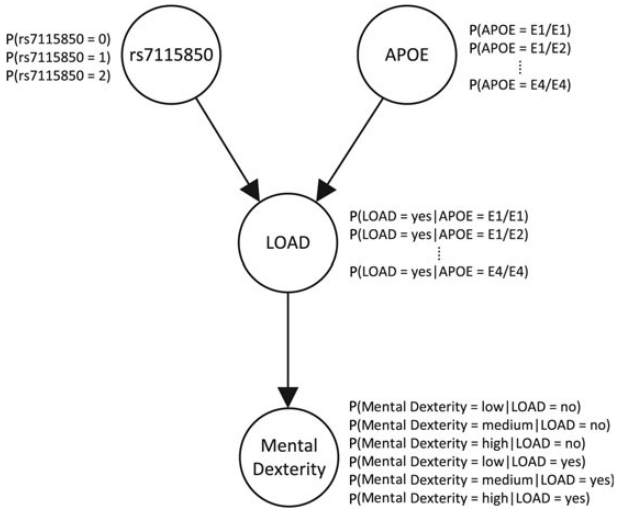

Figure 1 A Bayesian network modeling the situation that single nucleotide polymorphism (SNP) rs7115850 and the APOE gene cause late onset Alzheimer's disease (LOAD), and mental dexterity. SNPs have two alleles A and B; so the value of an SNP in a human being can be AA, AB, or BB, which we label 0, 1, and 2. The APOE gene has four alleles: E1, E2, E3, and E4. We have modeled three levels of mental dexterity.

Using a BN, we can determine conditional probabilities of interest with a BN inference algorithm.[45] The problem of doing inference in BNs is NP-hard.[55]

The task of learning a BN from data concerns learning both the parameters in a BN and the structure (called a DAG model). In a score-based structure learning approach, we assign a score to a DAG based on how well the DAG fits the data. Cooper and Herskovits[56] introduced the Bayesian score, which is the probability of the Data given the model G. A popular variation of the Bayesian score is the Bayesian Dirichlet equivalent uniform (BDeu) score,[57] which allows the user to specify priors for the conditional probability distributions using a single hyperparameter α, called the prior equivalent sample size. That score is as follows:

$$\text{score}_\alpha(G : \text{Data}) = P(\text{Data}|G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha/q_i)}{\Gamma(\alpha/q_i + \sum_{k=1}^{r_i} s_{ijk})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha/r_i q_i + s_{ijk})}{\Gamma(\alpha/r_i q_i)} \quad (1)$$

where r_i is the number of states of node X_i, q_i is the number of different instantiations of the parents of X_i, and s_{ijk} is the number of times in the data that X_i took its kth value when the parents of X_i had their jth instantiation.

Finding the DAG model that maximizes a Bayesian score is NP-hard.[58]

Bayesian networks classifiers

We present four different but related BN classifiers.

Naive Bayesian networks

Ideally, we would like to identify all the SNPs that predict a clinical disease outcome of interest, and have those SNPs be parents of the disease outcome in a BN. However, unless there are only a few predictors, we usually do not have sufficient data to learn such a network. For example, if all variables are binary and we have only 10 predictors, there are 1024 combinations of values of the predictors. An approach often taken to circumvent this dilemma is to make the predictors children of the outcome. Such a network is called an NB network, and when used for classification it is called an NB classifier.[39] Figure 2 shows the DAG model for an NB classifier when six SNPs are being used to predict disease D.

Suppose we have Data concerning the status of disease D and values of the six SNP predictors of D. To develop an NB classifier from these Data we learn the conditional probability distribution of each SNP S_i given D from the Data, and ascertain the prior probability of D in the population in which the classifier will be used. These probability distributions, along with the

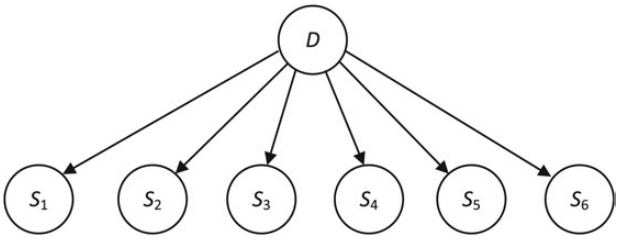

Figure 2 A directed acyclic graph (DAG) model for a naive Bayes (NB) classifier. The variable D represents disease status, and each variable S_i represents the state of a single nucleotide polymorphism (SNP).

DAG in figure 2, fully specify a BN. To use this network to predict disease status for a given patient with known values of the SNPs, we compute the following:

$$\begin{aligned} P(D|Data) &= KP(Data|D)P(D) \\ &= KP(S_1, S_2, S_3, S_4, S_5, S_6|D)P(D) \\ &= K[\prod_{i=1}^{6} P(S_i|D)]P(D), \end{aligned}$$

where K is a normalizing constant. The last equality above is because the SNPs are being modeled as independent of each other upon conditioning on their parent D. A naive BN classifier is a *probabilistic classifier* in that it provides the probability of disease status given Data.

Feature selection naive Bayes

When we have many possible SNP predictors, we do not know which of them are predictors of disease status. One alternative is to try to learn the predictors from the Data, and then use those predictors in an NB classifier. FSNB[40] uses this strategy. That is, it learns the predictors for a unique NB classifier, and then uses that unique classifier to perform inference. It learns the DAG model by starting with the model containing no SNPs. It then uses a greedy forward search that adds the SNP to the model that most increases the Bayesian score (equation 1) of the model. When no additional features increase the score, the search stops.

Model averaging naive Bayes

Another alternative is to consider all possible subsets of the set of all SNPs as possible predictors of the disease, compute the probability of the disease using an NB classifier containing each of the subsets, and then average over all the classifiers. Formally, we use the law of total probability as follows:

$$P(D|Data) = \sum_{M} P(D|Data, M)P(M|Data), \qquad (2)$$

where the sum is over all naive BN classifier models M containing subsets of the SNPs. This strategy is called MANB.[40 41] Since there are 2^n subsets of n SNPs, it is not possible to compute the sum in equation 2 by brute force. By exploiting the conditional independencies, it is possible to compute equation 2 in time that is linear in n.

Efficient Bayesian multivariate classifier

A problem with an NB network is that it makes strong conditional independence assumptions. That is, it assumes that the predictors are conditionally independent given the outcome, whereas often they are conditionally dependent given the outcome. The EBMC[42] ameliorates this difficulty. EBMC searches for predictors similar to FSNB, but does so in a more refined manner.

We describe the search algorithm in EBMC using an example that is adapted from Cooper *et al*.[42] As illustrated in figure 3, the algorithm starts by scoring all DAG models in which a single SNP is the parent of disease node D, using the BDeu score (equation 1). The model containing the highest scoring SNP is our initial model as shown in figure 3A, where we have labeled the SNP as S_1. We then determine which SNP when added as a parent of D to this 1-SNP model yields the highest scoring 2-SNP model. If that 2-SNP model has a higher score than our 1-SNP model, our new model becomes that 2-SNP model as depicted in figure 3B. We greedily keep adding SNPs to the model as long as we can increase the score. When no SNP increases the score further, we search for an SNP such that deleting the SNP increases the score, and we delete the SNP whose deletion increases the score the most. We keep deleting SNPs until no deletion increases the score.

Suppose our final model is the one in figure 3B. We then make the SNPs in the model children of D and create edges between them. The result is the model in figure 3C. This is the result we would obtain given the causal relationships in figure 1 if the loci we learned were SNP rs7115850 and the APOE gene. Next the search continues in the same manner, identifying additional SNPs that are parents of D. That is, we first identify the single SNP that when added as a parent of D to the model in figure 3C increases the score of that model the most. We again proceed with a greedy forward and backward search. Suppose the search yields two additional SNPs. We then have the model in figure 3D. In the same way as before, we make the new SNPs children of D and create edges between them. The result appears in figure 3E. The search repeatedly continues in this manner until we cannot increase the score further. The network produced by EBMC is called an augmented NB network.[59]

An algorithm for EBMC appears in the online supplementary material.

Other methods tested

We compared the BN-based methods to LR, which is a standard probabilistic binary classifier, the SVM,[43] which is a non-probabilistic machine learning binary classifier, Lasso,[20] which is regression-based and does shrinkage that allows a variable to be partly included in the model, and the neural network-based non-probabilistic binary classifier ELM.[43]

Evaluation methodology

We evaluated NB, MANB, FSNB, EBMC, LR, SVM, Lasso, and ELM using 110 simulated datasets, six semi-synthetic sets, and two real GWAS datasets. We used our own implementations of NB, MANB, FSNB, and EBMC, and the publicly available implementions of SVM at LIBLINEAR[60] and LIBSVM,[61] Lasso at lasso4j,[62] and ELM at ELM.[63] Our implementations are not yet publicly available. The SVM packages we used provides a linear kernel in LIBLINEAR and the radial basis function (RBF) kernel in LIBSVM. *Overfitting* occurs when a model describes the data well, but as a result describes the underlying relationships poorly. A model tries to avoid overfitting by employing *regularization*. The linear kernel uses a penalty parameter C to handle regularization. We used the following values of C in our studies: 2^{-5}, 2^{-1}, 2^{3}, 2^{7}, 2^{11}, and 2^{15}. The RBF kernel uses both the parameter C and a kernel parameter γ. We used the same values of C as those used for the linear kernel, and the following values of γ: −15, −11, −7, −3, 1, 5. We combined every value of C with every value of γ. In ELM we can specify the number of hidden nodes. We ran ELM with 10, 500, and 1000 hidden nodes. In the BN-based methods, we used the BDeu score with α=9. All experiments were run using a Dell PowerEdge R515 server which has an AMD Opteron 4276HE, 2.6 GHz, 8C, Turbo CORE, 8M L2/8M L3, 1600 MHz Max Mem single processor and an additional AMD Opteron 4276HE, 2.6 GHz, 8C, Turbo CORE, 8M L2/8M L3, 1600 MHz Max Mem processor.

Simulated datasets

Chen *et al*[64] generated datasets based on two 2-SNP epistatic interactions, two 3-SNP epistatic interactions, and one 5-SNP

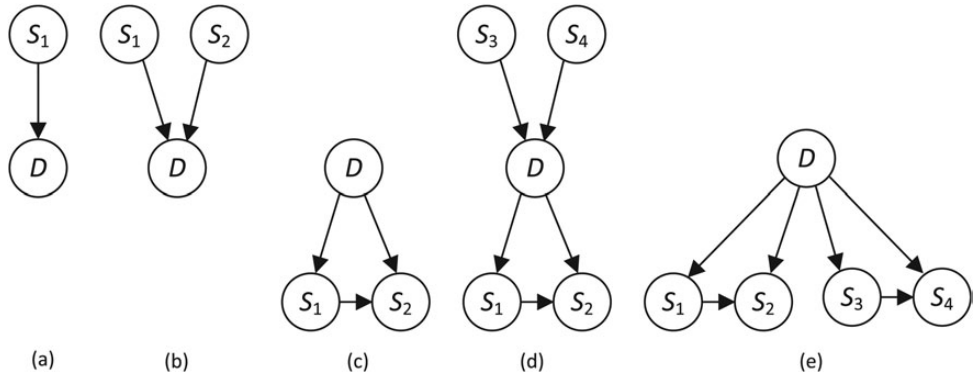

Figure 3 An example illustrating the efficient Bayesian multivariate classifier (EBMC) search. The variable D represents disease status, and each variable S_i represents the state of a single nucleotide polymorphism (SNP).

epistatic interaction, making a total of 15 predictive SNPs. Each dataset contains 1000 cases and 1000 controls. Three parameters were varied to create the interactions: (1) θ, which determined the penetrance; (2) β, which determined the minor allele frequency; and (3) l, which determined the linkage disequilibrium of the true causative SNPs with the observed SNPs. The effects of the interactions were combined using a Noisy-OR model.[45]

In our evaluation, we used the hundred 1000-SNP datasets and the ten 10 000-SNP datasets developed by Chen *et al*, with $\beta = 1$, $\beta = 1$, and l = null.

Real and semi-synthetic datasets

Reiman *et al*[65] developed a LOAD GWAS dataset that contained data on 312 260 SNPs and had records on 861 cases and 644 controls. We used this real GWAS dataset in our evaluation. Hunter *et al*[66] conducted a GWAS on 546 646 SNPs and breast cancer. The dataset consists of 1145 cases and 1142 controls. In addition to using this real GWAS dataset directly in our evaluation, we also used it to create semi-synthetic datasets. To create one of these datasets, we generated data on 15 predictive SNPs in the same way as described above for the simulated datasets. We then injected these data into the real breast cancer GWAS dataset, resulting in a semi-synthetic dataset. We developed six such semi-synthetic datasets.

RESULTS

Simulated datasets

In *k-fold cross-validation*, we divide the data into k partitions of the same size. For each partition j, we train using the data in the remaining $k-1$ partitions to yield a model, and we compute the error for each data item in partition j when applying this model. We evaluated the methods using fivefold cross-validation. In order to judge to what extent a method overfit the data, we also carried out *in-sample testing* in which we used the entire dataset to learn the model, and then used the learned model to do prediction for the entire dataset. Table 1 compares the average areas under the receiver operating characteristic curve (AUROC) for the fivefold cross-validation when the methods were used to analyze the simulated datasets. The best and worst results obtained for LIBLINEAR and LIBSVM over all values of the parameters are shown in table 1. Online supplementary table S1 compares the average AUROCs for the in-sample testing. Online supplementary table S2 shows the running times. Figure 4 plots the average AUROCs for the methods. In the fivefold cross-validation, LIBSVM performed the best on the 1000-SNP datasets, but its performance degraded on the 10 000-SNP datasets. EBMC performed best on the 10 000-SNP datasets, exhibiting the same performance as on the 1000-SNP datasets. Other than NB, the BN-based methods did not do much better in the in-sample testing than in the fivefold cross-validation, indicating that they do not tend to overfit the data. On the other hand, NB and the non-Bayesian methods performed better in the in-sample testing, indicate that these methods tend to overfit the data. As can be seen in online supplementary table S2, LIBSVM is substantially slower than the other methods. On average it took EBMC, LIBLINEAR best, and LIBSVM best 2.2 m, 4.77 s, and 0.44 h, respectively, to handle one 10 000-SNP dataset.

LIBSVM is a non-probabilistic classifier. For such classifiers it is worthwhile to investigate the actual true positive and false positive rates. LIBSVM achieved its average AUROC of 0.760 on the 1000-SNP datasets with an average true positive rate of 0.548 and an average false positive rate of 0.027. In applications where it is more important to avoid false positives, this performance would be good. On the other hand, if true positives are more important and some false positives can be tolerated, this would not be good performance.

Table 1 Average area under the receiver operating characteristic curves (AUROCs) when fivefold cross-validation is used to analyze the 100 simulated datasets containing 1000 single nucleotide polymorphisms (SNPs) and the 10 datasets containing 10 000 SNPs

Method	1000 SNPs	10 000 SNPs
EBMC	0.709	0.707
MANB	0.682	0.678
FSNB	0.682	0.679
NB	0.583	0.666
LR	0.553	0.516
LIBLINEAR best	0.593 ($C=2^{-5}$)	0.517 ($C=2^{-5}$)
LIBLINEAR worst	0.539 ($C=2^{11}$)	0.507 ($C=2^{11}$)
LIBSVM best	0.760 ($C=2^{3}$, $\gamma=-7$)	0.675 ($C=2^{3}$, $\gamma=-11$)
LIBSVM worst	0.250 ($C=2^{-3}$, $\gamma=-3$)	0.250 ($C=2^{-5}$, $\gamma=-7$)
Lasso	0.528	0.512
ELM 10	0.501	0.500
ELM 500	0.508	0.503
ELM 1000	0.509	0.503

EBMC, efficient Bayesian multivariate classifier; ELM, extreme learning machines; FSNB, feature selection naive Bayes; LR, logistic regression; MANB, model averaging naive Bayes; NB, naive Bayes.

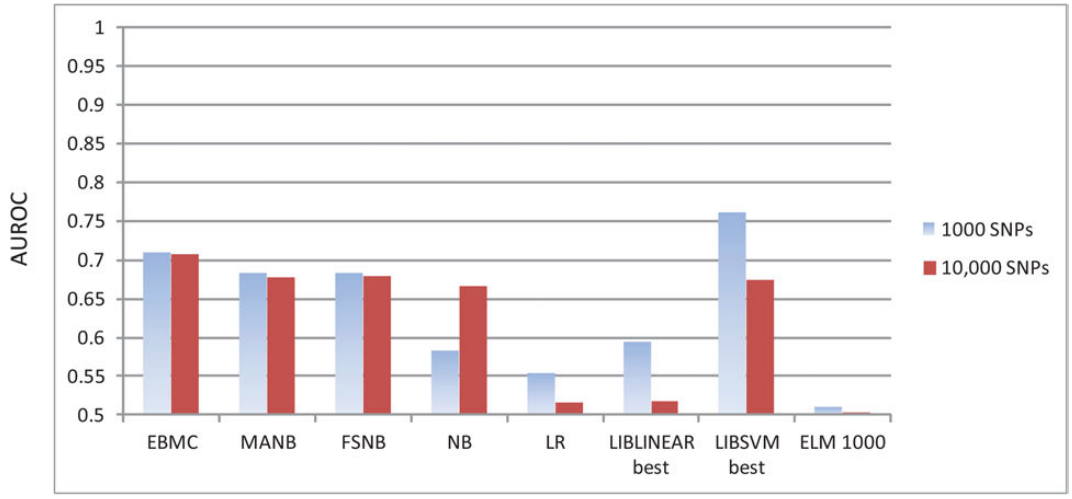

Figure 4 Average area under the receiver operating characteristic curves (AUROCs) for the methods when the 1000-single nucleotide polymorphism (SNP) and 10 000-SNP datasets are analyzed. EBMC, efficient Bayesian multivariate classifier; ELM, extreme learning machines; FSNB, feature selection naive Bayes; LR, logistic regression; MANB, model averaging naive Bayes; NB, naive Bayes; SNP, single nucleotide polymorphism.

Real and semi-synthetic datasets

We again evaluated the systems using fivefold cross-validation and in-sample testing. Lasso was not included in the evaluation because it could not handle so many SNPs. LIBSVM was also not included owing to memory problems; we could not get LIBSVM to run even when 60 GBs were allocated. Table 2 shows the AUROCs for the fivefold cross-validation when the methods analyze the real LOAD and breast cancer datasets, and the average AUROC when they analyze the semi-synthetic datasets. Online supplementary table S3 shows the in-sample testing results, while online supplementary table S4 shows the running times. Figure 5 plots the AUROCs for the methods when analyzing the LOAD dataset.

The fivefold cross-validation results for the LOAD dataset were similar to those for the simulated datasets. We investigated EBMC further when analyzing the LOAD dataset. Figure 6 shows the BN model learned by EBMC from the entire LOAD dataset. APOE is the strongest genetic risk factor for LOAD.[65] SNP rs7115850 is on the GAB2 gene, and previous research has indicated the APOE and GAB2 genes interact to affect LOAD.[65] The edge from SNP rs7115850 to APOE captures this interaction. SNP rs6784615 is on the NISCH gene; recent research has associated this gene by itself with LOAD.[67] Note that EBMC did not place an edge between this SNP and either of the other predictors. So these associations in the EBMC model are consistent with studies in the literature. The models obtained in the fivefold cross-validation studies were not as consistent with known biological information as the model in figure 6. They all contained the APOE gene, but none found the GAB2 gene. Model 1 contained the NISCH gene and APOE, model 3 contained only APOE, model 4 contained SNP rs7335085 and APOE, and model 5 contained SNP rs58766952 and APOE. We know of no previous research linking these latter two SNPs to LOAD. The fact that we learn from less data in the cross-validation studies may account for the inferior discovery. On the other hand, the fact that APOE is such a strong signal likely accounts for the fact that the performance is still good in these studies.

Table 2 Area under the receiver operating characteristic curves (AUROCs) when fivefold cross-validation is used to analyze the real genome-wide association (GWAS) datasets and the six semi-synthetic breast cancer datasets

Method	LOAD	Breast cancer	Avg. synthetic
EBMC	0.710	0.499	0.699
MANB	0.722	0.489	0.688
FSNB	0.692	0.499	0.690
NB	0.558	0.529	0.804
LR	0.542	0.504	0.516
LIBLINEAR best	0.607 ($C=2^{-5}$)	0.524 ($C=2^{7}$)	0.569 ($C=2^{-5}$)
LIBLINEAR worst	0.533 ($C=2^{15}$)	0.502 ($C=2^{-5}$)	0.501 ($C=2^{7}$)
ELM 10	0.490	0.503	0.497
ELM 500	0.533	0.481	0.491
ELM 1000	0.514	0.497	0.507

EBMC, efficient Bayesian multivariate classifier; ELM, extreme learning machines; FSNB, feature selection naive Bayes; LOAD, late onset Alzheimer's disease; LR, logistic regression; MANB, model averaging naive Bayes.

The fivefold cross-validation breast cancer results are interesting and surprising. Most of the methods found no signal at all. However, LIBLINEAR found a very weak signal, and NB found a stronger but still weak signal. We obtained 95% CIs for the AUROCs of LIBLINEAR best and NB equal to (0.500 to 0.548) and (0.505 to 0.553), respectively, so we can be reasonably confident that this is a faint signal and not noise. When analyzing this dataset using the Bayesian network posterior probability (BNPP) method, Jiang *et al*[68] found that no SNP had a posterior probability great than 0.01, but there were many SNPs with posterior probabilities close to 0.003. This indicates that there is no strong predictor in this dataset, but many possible weak ones. (Note that BRCA1 and BRCA2 are not included in this dataset because they are too rare to qualify. Furthermore, the study is in postmenopausal women, and these genes are known to be risk factors in premenopausal women.) So, many SNPs could be interacting to provide this weak signal. Such a result is consistent with the fact that studies have shown that thousands of genes may be associated with breast cancer.[69] Furthermore, when we inject the five interactions into the breast cancer dataset, NB does better than the other methods and also better than any of the results for the simulated datasets. NB does

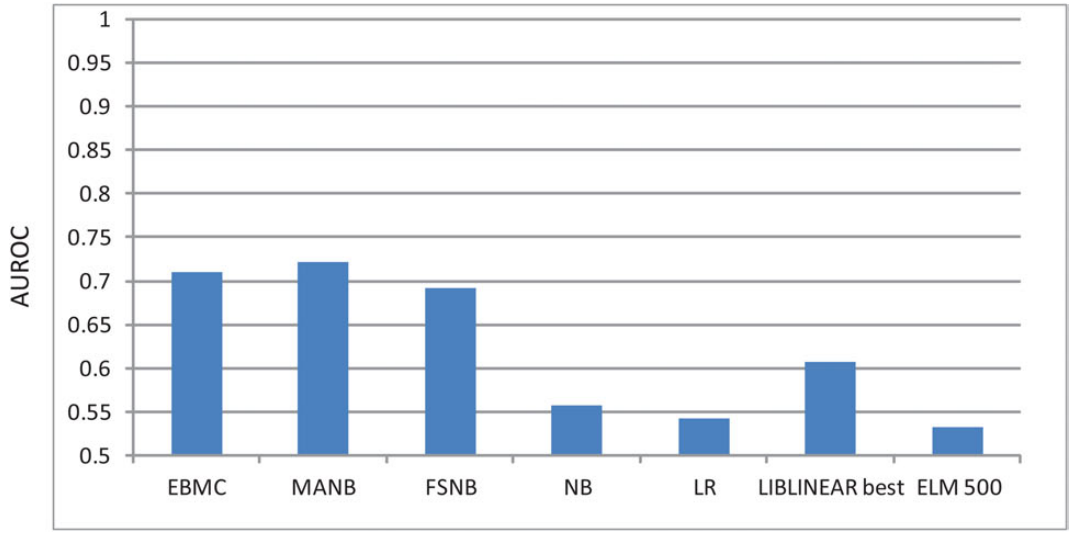

Figure 5 Area under the receiver operating characteristic curves (AUROCs) for various methods when the late onset Alzheimer's disease (LOAD) dataset is analyzed. EBMC, efficient Bayesian multivariate classifier; ELM, extreme learning machines; FSNB, feature selection naive Bayes; LR, logistic regression; MANB, model averaging naive Bayes; NB, naive Bayes.

prediction using all possible predictors. It seems that this strategy may work well in this situation because it combines the predictive value of the injected SNPs with the weak signal of the multitude of real predictors to achieve its predictive capability. Further analysis with additional real and real/combined datasets is needed to investigate this conjecture.

DISCUSSION

LIBSVM performed best on the 1000-SNP datasets. EBMC performed best on the 10 000-SNP datasets and slightly behind MANB on the LOAD dataset. In the studies on the breast cancer dataset, NB performed best. EBMC searches for good predictors and then uses them to carry out prediction, whereas NB simply uses all possible predictors. Our results indicate that the former approach might work better when there are several strong predictors, whereas the latter approach might work better when there are many weak predictors.

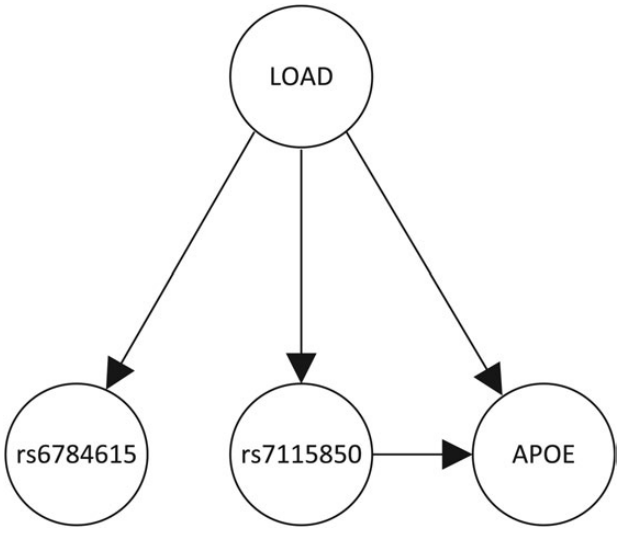

Figure 6 The Bayesian network model learned by the efficient Bayesian multivariate classifier (EBMC) from the entire late onset Alzheimer's disease dataset.

Using the same simulated datasets as those used in our analysis, Chen *et al*[64] compared the performance of eight methods for discovering epistatic interactions. All methods performed substantially worse on the 10 000-SNP datasets than on the 1000-SNP datasets. Since the 10 000-SNP datasets contain more non-predictors, it is more difficult to find the 15 actual predictors when analyzing these datasets. However, our prediction results using EBMC obtained AUROCs of around 0.70 for the 1000-SNP datasets, the 10 000-SNP datasets, and the semi-synthetic breast cancer datasets containing 546 646 SNPs. This result indicates that EBMC's predictive ability hardly degrades as we increase the number of SNPs, even if the discovery performance does degrade.

In future research, we can generate more simulated datasets with different numbers of interacting SNPs and different values of θ, β, and l. We can then further investigate the performance of LIBSVM, EBMC, and NB using various values of their tuning parameters. We can identify which method performs best in each situation.

We evaluated our methods using a real LOAD GWAS dataset which has a strong SNP signal, and a real breast cancer GWAS dataset which has a weak SNP signal. Future research can further investigate their performance using a real GWAS dataset in a domain believed to have a moderate SNP signal such as rheumatoid arthritis.[70]

CONCLUSIONS

Our investigations support the hypothesis that EBMC performs binary prediction using high-dimensional genomic datasets well. It exhibited the best overall performance of eight methods tested, although LIBSVM, MANB, and NB performed better in certain cases.

Contributors REN and XJ: jointly developed the structure and arguments for the manuscript and analyzed the data; XJ: conceived and designed the experiments; DX and BC: conducted the experiments; REN: wrote the first draft of the manuscript; GFC: developed the EBMC algorithm and wrote the version that appears in the online supplementary material; XL: contributed to the writing of the manuscript; GFC and XL: agreed with manuscript results and conclusions and made critical revisions. All authors reviewed and approved the final manuscript.

Funding Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R00LM010822.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Brookes AJ. The essence of SNPs. *Gene* 1999;234:177–86.
2 Ng PC, Kirness EF. Whole genome sequencing. *Methods Mol Biol* 2010;628:215–26.
3 Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. *Annu Rev Med* 2009;60:443–56.
4 Herbert A, Gerry NP, McQueen MB. A common genetic variant is associated with adult and childhood obesity. *J Comput Biol* 2006;312:279–384.
5 Spinola M, Meyer P, Kammerer S, *et al*. Association of the PDCD5 locus with long cancer risk and prognosis in smokers. *Am J Hum Genet* 2001;55:27–46.
6 Lambert JC, Heath S, Even G, *et al*. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* 2009;41:1094–9.
7 Galvin A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet* 2010;26:132–41.
8 Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases and complex traits. *Nature* 2009;461:747–53.
9 Mahr B. Personal genomics: the case of missing heritability. *Nature* 2008;456:18–21.
10 Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 2010;26:445–55.
11 Nagel RI. Epistasis and the genetics of human diseases. *C R Biol* 2005;328:606–15.
12 Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003;19:376–82.
13 Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. *Ann Med* 2002;34:88–95.
14 Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–47.
15 Cho YM, Ritchie MD, Moore JH, *et al*. Multifactor dimensionality reduction reveals a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia* 2004;47:549–54.
16 Jiang X, Neapolitan RE, Barmada MM, *et al*. Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics* 2011;12:1471–2105.
17 Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 2005;28:157–70.
18 Agresti A. *Categorical data analysis*. 2nd edn. New York: Wiley, 2007.
19 Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008;9:30–50.
20 Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 1998;20:33–61.
21 Wu TT, Chen YF, Hastie T, *et al*. Genome-wide association analysis by lasso penalized logistic regression. *Genome Anal* 2009;25:714–21.
22 Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;37:413–17.
23 Moore JH, Gilbert JC, Tsai CT, *et al*. A flexible computational framework for detecting characterizing and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* 2006;241:252–61.
24 Jakulin A, Bratko I. Testing the significance of attribute interactions. *Proceedings of the 21st International Conference on Machine Learning (ICML-2004)*. Banff, Canada, 2004.
25 Yang C, He Z, Wan X, *et al*. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* 2009;25:504–11.
26 Zhang X, Pan F, Xie Y, *et al*. COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study. *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. Tuscon, Arizona, 2009.
27 Wongseree W, Assawamakin A, Piroonratana T, *et al*. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC Bioinformatics* 2009;10:294.
28 Moore JH, White BC. Tuning ReliefF for genome-wide genetic analysis. In: Marchiori E, Moore JH, Rajapakee JC, eds. *Proceedings of EvoBIO 2007*, Berlin: Springer-Verlag, 2007:166–75.
29 Epstein MJ, Haake P. Very large scale ReliefF for genome-wide association analysis. *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Sun Valley, Idaho, 2008.
30 Meng Y, Yang Q, Cuenco KT, *et al*. Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and Bayesian networks. *BMC Proc* 2007;1(Suppl 1):S56.
31 Wan X, Yang C, Yang Q, *et al*. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* 2007;26:30–7.
32 Logsdon BA, Hoffman GE, Mezey JG. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* 2010;11:58.
33 Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case control studies. *Nat Genet* 2007;39:1167–73.
34 Verzilli CJ, Stallard N, Whittaker JC. Bayesian graphical models for genomewide association studies. *Am J Hum Genet* 2006;79:100–12.
35 Miller DJ, Zhang Y, Yu G, *et al*. An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics* 2009;25:2478–85.
36 Han B, Park M, Chen X. A Markov blanket-based method for detecting causal SNPs in GWAS. *Proceeding of IEEE International Conference on Bioinformatics and Biomedicine*. Washington, DC, 2009.
37 Li J, Horstman B, Chen Y. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics* 2011;27:222–9.
38 Jiang X, Neapolitan RE, Barmada MM, *et al*. A fast algorithm for learning epistatic genomics relationships. *AMIA Annu Symp Proc* 2010;2010:341–5.
39 Jensen FV, Neilsen TD. *Bayesian networks and decision graphs*. New York: Springer-Verlag, 2007.
40 Wei W, Visweswaran S, Gooper GF. The application of naïve Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J Am Med Inform Assoc* 2011;18:370–5.
41 Dash D, Cooper G. Model averaging for prediction with discrete Bayesian networks. *J Mach Learn Res* 2004;5:1177e203.
42 Cooper GF, Yeomans PH, Visweswaren S, *et al*. An efficient Bayesian method for predicting clinical outcomes from genome-wide data. *Proceedings of AMIA 2010*. Washington, DC, 2010.
43 Cortes C, Vapnik VN. Support-vector networks. *Mach Learn* 1995;20:273–97.
44 Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. *Neurocomputing* 2006;70:489–501.
45 Neapolitan RE. *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall, 2004.
46 Neapolitan RE. *Probabilistic reasoning in expert systems*. New York, NY: Wiley, 1989.
47 Pearl J. *Probabilistic reasoning in intelligent systems*. Burlington, MA: Morgan Kaufmann, 1988.
48 Korb K, Nicholson A. *Bayesian artificial intelligence*. Boca Raton, FL: CRC Press, 2010.
49 Segal E, Pe'er D, Regev A, *et al*. Learning module networks. *J Mach Learn Res* 2005;6:557–88.
50 Friedman N, Linial M, Nachman I, *et al*. Using Bayesian networks to analyze expression data. *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*. Tokyo, Japan, 2005.
51 Fishelson M, Geiger D. Optimizing exact genetic linkage computation. *J Comput Biol* 2004;11:263–75.
52 Friedman N, Koller K. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Mach Learn* 2003;20:201–10.
53 Friedman N, Ninio M, Pe'er I, *et al*. A structural EM algorithm for phylogenetic inference. *J Comput Biol* 2002;9:331–53.
54 Fishelson M, Geiger D. Exact genetic linkage computations for general pedigrees. *Bioinformatics* 2002;18:S189–98.
55 Cooper GF. The computational complexity of probabilistic inference using Bayesian belief networks. *J Artif Intell* 1990;42:393–405.
56 Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn* 1992;9:309–47.
57 Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. Technical Report MSR-TR-94-09. Microsoft Research, 1995.
58 Chickering M. Learning Bayesian networks is NP-complete. In: Fisher D, Lenz H, eds. *Learning from data: artificial intelligence and statistics V*. New York, NY: Springer-Verlag, 1996:121–31.
59 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn* 1997;29:131–63.
60 http://www.csie.ntu.edu.tw/~cjlin/liblinear/
61 http://www.csie.ntu.edu.tw/~cjlin/libsvm/
62 https://code.google.com/p/lasso4j/downloads/detail?name=lasso4j-1.0.jar
63 http://www.ntu.edu.sg/home/egbhuang/source_codes/ELM.zip
64 Chen L, Yu G, Langefeld CD, *et al*. Comparative analysis of methods for detecting interacting loci. *BMC Genomics* 2011;12:344.

65 Reiman EM, Webster JA, Myers AJ, *et al*. GAB2 alleles modify Alzheimer's risk in APOE carriers. *Neuron* 2007;54:713–20.
66 Hunter DJ, Kraft P, Jacobs KB, *et al*. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007;39:870–4.
67 Briones N, Dinu V. Data mining of high density genomic variant data for prediction of Alzheimer's disease risk. *BMC Med Genet* 2012;13:7.
68 Jiang X, Barmada MM, Cooper GF, *et al*. A Bayesian method for evaluating and discovering disease loci associations. *PLoS ONE* 2011;6:e22075.
69 Nev R, Chin K, Fridlyand J, *et al*. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 2007;10:515–27.
70 Stahl EA, Raychaudhuri S, Remmers EF, *et al*. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 2010;42:508–14.