Advances in Neural Information Processing 9 (N485ff96), editors orksMozer, M.I. Jordan and T. Petsch

A BC Cancer Plotted22C West 10th Ave, Epidemiology statistical techniques modeling the relationship betwtobacco and alco

Pierre

tap@comp.vuw.ac.nz

DeptofChemicalerUniversityofBritishColumer8216MainMall

effects of predictors are linear (or categorical) and additive on the log-odds scale.
Variablesnotapplicabletoaparticularsub ariableswerezero,e.g.,numberofyearsof
An important virtue of logistic regression is that the relationships identified in the data can be interpreted and explained in simple terms, such as "the odds of developing lung cancer for males who smoke between 20 and 29 cigarettes per day increased by a factor of 11.5 ove

their simplicity, it is difficult to use these models to discover unanticipated complex
ducted by the Division of Epidemiology and Cancer Prevention at the BC Cancer
relationships, i.e., non-linearities in the effect of a predictor or interactions between
time tobacco and alcohol use, and lifetime employment history for each subject.
predictors. Interactions and non-linearities can of course be introduced into regressions, but must be pre-specifi

are only a few vorthereareapriorireasonstotestforparticulareffects.

Neural networks have the potential to automatically discover complex ariablesThere has been much interest in u

witness the recent series of articles in *The Lancet*, e.g., Wyatt [1995] and Baxt [1995]. However, there are not yet sufficient comparisons or theory to come to firm conclusions about the utility of neural networks in biomedical data analysis. To date, comparison studies, e.g, those by Michie, Spiegelhalter, and Taylor [1994], BurkRosen,andGoodman[1995],andLippmann,Lee,andShahian[1995],have had mixed results, and Jefferson et al's [1995] complaint that many "successful" applications of neural networks are not compared against standard techniques appears to be justified. e,intentofthispaperistocontributetothebodyofuseful comparisons by reporting a study of various neural-network and statistical modeling techniques applied to an epidemiological data analysis problem.

2 The data

Theoriginaldatasetconsistedofinformationon15,463sub Thefromastudycon-

Agency. In this study, detailed questionnaire reported personal information, life-

The subjects were cancer patients in BC with diagnosis dates between 1983 and 1989, as ascertained by the population-based registry at the BC Cancer Agency. jectsdifferenttobaccoandalcoholhabitswereincluded: cigarette (C), cigar (G), Sixpipe (P) smoking, and beer (B),

pinthispaperusedupto27predictorvariables: ortedatfirstdiagnosis(AGE), and26vagerelatedtoalcoholandtobaccoconsumption. These included four variables for each habit: total years of consumption (CYR etc), consumption pdaorweek(CDAY, BWK etc), yea

vindicatedwhetherthesubject ever smoked tobacco or drank alcohol. All the binary variableswerenon-linear(threshold)transformsoftheothervariables.

smokingforanon-smoker,oryearssincequittingforasmokerwhodidnotquit.

Of the 15,463 records, 5901 had missing information in some of the fields related to tobacco or alcohol use. These were not used, as there are no simple methojectfor dealing with missing da

to be used in the modeling experimenjectThere were 28 binary outcomes: the 28 sites at which a subject could

varied, e.g., for LUNGSQ (Lung Squamous) there were 694 cases among the Therecords, for ORAL (Oral Cavity an

All sites were modeled individually using carefully selected subjects as controls.

This is common practice in cancer epidemiology studies, due to the difficulty of

Logistic regression: The FULL model used the full set of predictor visit.ulaquadratictermforage: AGESQ.

collecting an unbiased sample of non-cancer subjects for controls. Subjects with cancers at a site suspected of being related to tobacco usage were not used as

AdditiveModelsorGAMs)tocompletelyflexible(classificationtrees

andMMY(Multiple-Myeloma),andresultedinbetween2959and3694

andneuralnetworks). Each site was modeled independently, using the log likeli-

controls for each site. For example, the model for LUNGSQ (lung squamous cell)

lymphoma)wasfittedusingsubcancerwithLUNGSQasthepositiveoutcomes(694jectsandsubjects all of whose sites we

controls. This eliminated subjects with any sites other than COLON, RECTUM, (Melanoma), NMSK (Non-melanc

or MMY as negative outcomes (3694 controls).

parisons to test this, one of non-flexible models (ONLYLIN-STEP) also did not use any

3 Statistical methofiectsA number of different types of statistical methods were used to model the

These ranged from the non-flexible (logistic regression) through partially flexible

hood of the data under the binomial distribution as the fitting criterion. (Generalizedofmodeling,exceptforthene

theS-plusstatisticalsoftwarepackAll[StatSci1995].

For several methods, we used Breiman's [1996] *bagging* technique to control overfitting. To "bag" a model, one fits a set of models independently on bootstrap samples. The bagged prediction is then the average of the predictions of the models in the set. Breiman suggests that bagging will give superior predictions for unstable moage(suchasstepwiseselection,prunedtrees,andneuralnetworks).

Preliminary analysis revealed that the predictive power of non-flexible models delsbe improved by including non

and the binary indicator variables SMOKE, DRINK, CSMOKE, etc. Flexible models

should be able to discover useful non-linear transforms for themselves and so these derived vwerenotincludedintheflexiblemodels. In order to allow com-

ariablesthesederivedvariables.

Null model: (NULL) The predictions of the null model are just the frequency of the outcome in the training

Stepwise logistic regression: includingnumberofstepwiseregressionswerefitted,differing in the set of vAconsidered. 10-foldcrossvalidation

wOutcome-balancedusedtoselectthemodelsizegivingbestgeneralization. The models were as [illegible] best numbforof hidden units were trained on the en

spline of degree 2 for each parameter, and stopped when the AIC statistic could not reduced any further. Two stepwise GAM models were fitted: GAM-FULL used the full set of variables, while GAM-CIG used the cigarette vTheandAGE.

Classification trees: [Breiman et al. 1984] The same cross-validation procedure as used with stepwise regression was used to select the best size for TREE, using the implementation in S-plus, and the function shrink.tree() for pruning. A bagged version with 50 replications, TREE-BAGGED, was also used. After constructing a tree for the data in a replication, it was pruned to perform optimally on the ariablesnotincludedinthatreplication.

Ordinaryneuralnetworks: dataneuralnetworkmodelshadasinglehiddenlayer oftanhfunctionsandasmallweightpenalty(0.01)topreventparametersgoingto infinity. Theconjugate-gradientprocedurewasusedtooptimizeweights. For the NN-ORD-H2 model, which had no control on complexity, a network with two hidden units was trained three times from different small random starting weights. Of these three, the one with best performance on the training data was selected as "the model". The NN-ORD-HCV used common method for controlling overfitting in neural networks: 10-fold CV for selecting the optimal number of hidden units. Three random starting points for each partition were used calculate the average generalization error

haAthelowesttrainingerrorwaschosen.

Bagged neural networks with early stopping: Bagging and early stopping (terminating training before reaching a minimum on training set error in order vingprevent overfitting) work n

early stopping, training is fast enough to make bagging practical. 100 networks with two hidden units were trained on separate bootstrap replications, and the bving50(bytheirperformanceontheomittedexamples)wereincludedinthefinal baggedmodel,NN-ESTOP-BAGGED.For comparison purposes, the mean individual pestofthesеearly-stoppednetworksisreportedasNN-ESTOP-AVG.

Neural networks with Bayesian regularization: MacKay's [1995] Bayesian

dom starts for networks with 1, 2, 3 or 4 hidden units and three different sets of regularization (penalty) parameters were used, giving a total of 36 networks for each site. The three possibilities for regularization parameters were: (a) three penalty parameters – one for each of input to hidden, bias to hidden, and hidden to output; (b) partial Automatic Relevance Determination (ARD) [MacKay 1995] with seven penalty parameters controlling the input to hidden weights – one for each habit and one for AGE; and (c) full ARD, with one penalty parameter for each of the 19 inputs. erformance"evidence"foreachnetworkwasevaluated and the best 18 networks wTheselected for the equally-w

wThethesinglenetworkwiththemaximumevidence.

4 Results and Discussion

Models were compared based on their performance on the held-out test data, asas to avoid overfitting bias in

asandGoodman[1995]),weusedthetest-setdevianceasitseemsRosento compare models using the same criterion a (i.e., sites), and some models did better on some sites and worse on others. We rep some overall performance figures and some pairwise comparisons of models.

bthanthoseofONLYLIN-STEP in 99.82% of resampled test sets (out of 10,000),

while the predictions of NN-BAYES-CMTT beat FULL-STEP in 93.75% of replications and FULL-STEP beat ONLYLIN-STEP in 98.48% of replications.

These results demonstrate that good control on overfitting is essential for this task. Ordinary neural networks with no control on overfitting do worse than guessing (i.e., the null model). Even when the number of hidden units is chosen by cross-validation, the performance is still worse than a simple two-variable stepwise logistic regression (CYR-AGE-STEP). The inadequacy of the simple AIC-based stepwise procedure for choosing the complexity of GAMs is illustrated by the poor performance of etterGAM-FULL model (the more

The effective methods for controlling overfitting were bagging and Bayesian regularization. Bagging improved the performance of trees and early-stopped neural netwto good levels. Bayesian regularization worked very well with neural networks and with ridge regression. Furthermore, examination of the performance of individual networks indicates that networks with fine-grained ARD were frequently superior to those with coarser control on regularization.

5 Artificial sites with complex relationships

The very minor improvement achieved by neural networks and trees over logistic models provokes the following question: are complex relationships are really relatively unimportant in this data, or is the strong control on overfitting preventing identification of complex relationships? In order to answer this question, we created six artificial "sites" for the subjects. These were designed to have very similar properties to the real sites, while possessing non-linear effects and interactions.

The risk models for the artificial sites possessed a underlying trend equal to half that of a good logistic model for LUNGSQ, and one of three more complex effects: FREQ, a frequent non-linear (threshold) effect (BWK > 1) affecting 4,334 of the orkssubjects; RARE, a rare thresh

INTER, an interaction ($\cdot$ GYR) affecting 482 subjects. For three of the artificial sites the complex effect was weak (LO), and for the other three it was strong (HI). For each subject and each artificial site, a random choice as to whether that subject was a positive case for that site was made, based on probability given by the model for the artificial site. Models were fitted to these sites in the same way as to other sites and only subBffRwithoutcanceratasmokingrelatedsitewereusedascontrols.

the model RARE-TRUE contained the continuous VAGE,AGESQ,CDAY, CYR,

and CYQUIT, and the binary variablesSMOKEandBWK>10.

Figure2showsperformanceontheartificialdata. The neural networks and bagged trees were very effective at detecting non-linearities and interactions. ariablespmancewasatthesamelevelastheapp

of simple models lacking the ability to fit the complexities (e.g., FULL-STEP) was Theirw

6 Conclusions

For predicting the risk of cancer in our data, neural networks with Bayesian estimation of regularization parameters to control overfitting performed consistently but only slightly better than logistic regression models. This appeared to be due to the lack of complex relationships in the data: on artificial data with complex relationships they performed markedly better than logistic models. Good control of overfitting is essential for this task, as shown by the poor performance of neural networks with the number of hidden units chosen by cross-validation.

Given their ability to not overfit while still identifying complex relationships we expect that neural networks could prove useful in epidemiological data-analysis by providing a method for checking that a simple statistical model is not missing important complex relationships.

Acknowledgments

This research was funded by grants from the Workers Compensation Board of British Columbia, NSERC, and IRIS, and conducted at the BC Cancer Agency.

References

ﬁaxtﬁ ﬁﬁ ﬁﬁ 1ﬁﬁ5ﬁ Application of artificial neural networﬁs to clinical medicineﬁ ﬁancetﬁ 346ﬁ1135

ﬁreimanﬁ ﬁﬁ 1ﬁﬁ6ﬁ ﬁaﬁﬁinﬁ predictorsﬁ ﬁachine ﬁearninﬁﬁ 24