Bayesian network modeling metrics of performance and uncertainty

by: Bruce G. Marcot
version 5 December 2012; updated 13 August 2014

originally produced as an adjunct to: Marcot, B. G. 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. Ecological Modelling 230:50-62.

PPD = posterior probability distribution

Model sensitivity analysis			
variance reduction	$VR = V(Q) - V(Q\|F)$, where $V(Q) = \sum_q P(q)[X_q - E(Q)]^2$, $V(Q\|F) = \sum_q P(q\|f)[X_q - E(Q\|f)]^2$, $E(Q) = \sum_q P(q)X_q$, where X_q is the numeric real value of state q, $E(Q)$ is the expected real value of Q before applying new findings, $E(Q\|f)$ is the expected real value of Q after applying new findings *f* for variable F, and $V(Q)$ is the variance of the real value of Q before any new findings	[0,infinity], the greater the value, the more sensitive is a resultant node Q to the node in question F; used with continuous variables	Marcot et al. 2006, Marcot 2012
entropy reduction	Entropy reduction, *I*, is calculated as the expected reduction in mutual information of Q from a finding for variable F, calculated as $I = H(Q) - H(Q\|F) = \sum_q \sum_f \frac{P(q,f) log_2 [P(q,f)\}}{P(q)P(f)}$ where $H(Q)$ is the entropy of Q before any new findings, $H(Q\|F)$ is the entropy of Q after new findings from variable F, and Q is measured in information bits	[0,infinity] , the greater the value, the more sensitive is a resultant node Q to the node in question F; used with discrete variables	Marcot et al. 2006, Marcot 2012
case file simulation	generate a large number of simulated data sets and analyzing the covariation between values of input variables and PPDs	not a scalar value; the higher the covariation, the greater the sensitivity	Marcot et al. 2006, Thogmartin 2010

Scenario analysis			
influence runs	evaluating effects on PPDs from selected input variables set to best- or worst-case scenario values	not a scalar value; the greater the deviation from the normative PPD, the greater is the influence	Marcot et al. 2012

Model complexity			
number of model variables	count of number of network nodes	[0,infinity]; the higher the value, the more complex the model	Marcot 2012
number of model links	count of number of network node connections	[0,infinity] ; the higher the value, the more complex the model	Marcot 2012

number of model node states	count of all states among all nodes	[0,infinity] ; the higher the value, the more complex the model	Marcot 2012
number of conditional probabilities	$\sum_{i=1}^{V}\left[S\prod_{j=1}^{n}P_j\right]$, where S = no. states of the child node, P_j = no. of states of the j^{th} parent node, for *n* parent nodes, among all *V* nodes in the model.	[0,infinity] ; the higher the value, the more complex the model	Marcot 2012
number of node cliques	count of all network cliques in the model	[0,infinity] ; the higher the value, the more complex the model	Marcot 2012

Prediction performance			
confusion matrix	numbers of false positives (Type I error, rejecting a true hypothesis), false negatives (Type II error, failing to reject a false hypothesis), and their sum	[0,n], n = total number of test cases; or [0,1] if put on a proportional basis; the higher the values, the greater is the classification error rate	Marcot 2012
covariate-weighted confusion error rate	confusion matrix Type I, Type II, and total error rates, times the number of covariates (nodes) in the model	[0,nc], n = total number of test cases, c = number of covariates in the model; or [0,1] if put on a proportional basis; the higher the values, the greater is the classification error rate	Marcot 2012
conditional probability-weighted confusion error rate	confusion matrix Type I, Type II, and total error rates, times the number of conditional probabilities in the model	[0,np], n = total number of test cases, p = number of conditional probabilities in the model; or [0,1] if put on a proportional basis; the higher the values, the greater is the classification error rate	Marcot 2012
AUC under ROC	area under the receiver operating characteristic (ROC) curve, which plots the percent true positives (“sensitivity”) as a function of their complement, percent false positives (“1-specificity”)	[0,1], where 1 denotes no error, 0.5 denotes totally random models, and < 0.5 denotes models that more often provide wrong predictions	Dlamini 2010, Hand 1997
k-fold cross-validation	one randomizes the case file set; sequentially numbers the resulting cases; extracts the first $1/k^{th}$ of the cases in sequence; parameterizes the model with the remaining $[1 – 1/k]$ cases; and then tests that model against the first $1/k^{th}$ cases left out, recording confusion error rates of model predication. Next, the second $1/k^{th}$ set of cases are extracted from the full case file set, and the procedure is repeated until all k case subsets have been used. The resulting *k* confusion tables are then averaged for overall model performance.	see above under confusion matrix	Boyce et al. 2002, Cheng and Greiner n.d.; also see Cawley and Talbot 2007 for the “leave-one-out cross-validation procedure”

spherical payoff	$SP = MOAC \cdot \frac{P_c}{\sqrt{\sum_{j=1}^{n} P_j^2}}$, where *MOAC* = mean probability value of a given state averaged over all cases, P_c = the predicted probability of the correct state, P_j = the predicted probability of state *j*, and *n* = total number of states	[0,1], 1 denotes best model performance	B. Boerlage, pers. comm., Marcot 2012
Schwarz' Bayesian information criterion	$BIC = -2 \cdot \ln(ML) + k \cdot \ln(n)$, where *ML* = maximum likelihood value, *k* = number of parameters in the model, and *n* = number of observations; then subtract the lowest BIC value among all alternative model forms being compared from the BIC value of each alternative model	the smallest difference (ΔBIC) denotes the best-performing and most parsimonious model, that is, the model that best balances model error and dimension	Schwarz 1978
true skill statistic (Hanssen-Kuiper discriminant or skill score)	for a 2-state outcome, $TSS = \frac{(ad-bc)}{(a+c)(b+d)}$, where a = true positives, b = Type II error (false positives), c = Type I errors (false negatives), and d = true negatives, all represented either as absolute or relative frequencies	useful only with 2x2 confusion matrices; [-1,1], where 1 represents a perfectly performing model with no error, 0 a model with totally random error, and -1 a model with total error	Allouche et al. 2006, Mouton et al. 2010
Cohen's kappa	the difference between correct observations and expected outcomes, divided by the complement of expected outcomes; Kappa is calculated as the difference between correct observations *O* and expected outcomes *E*, divided by the complement of expected, or $\kappa = \frac{O-E}{1-E}$, where $O = \frac{a+d}{N}$, $E = \frac{(a+b)(a+c)+(c+d)(b+d)}{N^2}$, and $N = (a+b+c+d)$.	[0,1], with 1 being perfect classification	Gutzwiller and Flather 2011, Zarnetske et al. 2007
logarithmic loss	LL = MOAC [- log (Pc)], where *MOAC* = mean probability value of a given state averaged over all cases, P_c = the predicted probability of the correct state	[0,infinity], 0 = best performance	Dlamini 2010, Norsys \1
quadratic loss (Brier score)	QL = MOAC [1 - 2 * Pc + sum[j=1 to n] (Pj ^ 2)], where *MOAC* = mean probability value of a given state averaged over all cases, P_c = the predicted probability of the correct state, P_j = the predicted probability of state *j*, and *n* = total number of states	[0,2], 0 = best performance	Norsys \1

Uncertainty of posterior probability distribution			
Bayesian credible interval	An *X*% Bayesian credible interval of some PPD of an ordinal or continuous scale variable (but not a categorical variable) refers to state-wise probabilities when *X/2*% is excluded from the lowest and highest outcome states. Put another way, it is the interval determined for the expected value over replicate calculations based on uncertainty distributions of the input variables, not for the PPD of a given instance of input values.	(not a scalar value)	Bolstad 2007, Curran 2005
posterior probability certainty index (PPCI)	PPDs which consist of p_i probability values among *N* number of states, where p_i ranges [0,1] and $\sum_{i=1}^{N} p_i = 1.0$. PPCI is calculated as (1-*J'*), where *J'* = H'/H'_{max}, $H' = -\sum_{i=1}^{N} p_i\, L$, where $L = \begin{cases} \ln(p_i), & p_i > 0 \\ 0, & p_i = 0 \end{cases}$, and $H'_{max} = ln(N)$. *J'* normalizes the metric proportional to *N*, so that the degree of certainty of PPDs can be compared among outcomes with different numbers of states *N*.	PPCI ranges [0,1] with higher values denoting greater certainty (greater loading of outcome probabilities into fewer outcome states). Models with higher PPCI values of their PPDs denote greater certainty in outcome predictions.	Marcot 2012
certainty envelope	$PPCI_{MIN}$ is calculated as $H' = -\left\{\sum_{i=1}^{j} P_i L + \sum_{i=j+1}^{N} \left[\left(\frac{1-m}{N-j}\right)_i \ln\left(\frac{1-m}{N-j}\right)_i\right]\right\}$ and $PPCI_{MAX}$ is calculated as $H' = -\left\{\sum_{i=1}^{j} P_i L + (1-m)\ln(1-m)\right\}$ where L is defined above.	For a given PPD with a specified probability of a given state or set of *j* states, $PPCI_{MIN} \leq [PPCI \mid P(j)] \leq PPCI_{MAX}$. To best compare PPCI values among competing models particularly with different total numbers of states *N* or different numbers of specified state values *j*, the range [$PPCI_{MIN}$, $PPCI_{MAX}$] can itself be normalized to [0,1], and the relative position of a given value of [PPCI | P(j)] within this range can be calculated by simple linear interpolation. Thus, the interpolated value of [PPCI | P(j)] represents the proportion (or percentage) of total possible certainty for a given outcome state(s) *j*.	Marcot 2012
Gini coefficient	calculated as the area under the Lorenz curve, which, applied to BN modeling, is the cumulative probability among outcome states rank-ordered by decreasing values of their individual probabilities; also see Marcot 2012 for a normalizing correction factor	2x the Gini coefficient ranges [0,1), where 0 represents a uniform probability distribution (complete uncertainty) and 1 represents a distribution with one state at 100% probability and all other states at 0% (complete certainty).	Marcot 2012

\1 http://www.norsys.com/WebHelp/NETICA/X_Scoring_Rule_Results.htm

References Cited

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43(6):1223-1232.

Bolstad, W. M. 2007. Introduction to Bayesian statistics. Second edition. Wiley-Interscience, New York. 464 pp.

Boyce, M. S., P. R. Vernier, S. E. Nielsen, and F. K. A. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling 157:281-300.

Cawley, G. C., and N. L. C. Talbot. 2007. Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. Journal of Machine Learning Research 8:841-861.

Cheng, J., and R. Greiner. n.d. Learning Bayesian belief network classifiers: algorithms and system. Department of Computing Science, University of Alberta. Edmonton, Alberta, Canada. 10 pp.

Curran, J. M. 2005. An introduction to Bayesian credible intervals for sampling error in DNA profiles. Law, Probability and Risk 4:115-126.

Dlamini, W. M. 2010. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland Environmental Modelling & Software 25(2):199-208.

Gutzwiller, K. J., and C. H. Flather. 2011. Wetland features and landscape context predict the risk of wetland habitat loss. Ecological Applications 21:968-982.

Hand, D. 1997. Construction and assessment of classification rules. John Wiley & Sons, New York. 232 pp.

Marcot, B. G., P. A. Hohenlohe, S. Morey, R. Holmes, R. Molina, M. Turley, M. Huff, and J. Laurence. 2006. Characterizing species at risk II: using Bayesian belief networks as decision support tools to determine species conservation categories under the Northwest Forest Plan. Ecology and Society 11(2):12. [online] URL: http://www.ecologyandsociety.org/vol11/iss2/art12/.

Marcot, B. G. 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. Ecological Modelling 230:50-62.

Mouton, A. M., B. De Baets, and P. L. M. Goethals. 2010. Ecological relevance of performance criteria for species distribution models. Ecological Modelling 221(16):1995-2002.

Schwarz, G. 1978. Estimating the dimension of a model. The Annals of Statistics 6(2):461-464.

Thogmartin, W. E. 2010. Sensitivity analysis of North American bird population estimates. Ecological Modelling 221(2):173-177.

Zarnetske, P. L., T. C. Edwards, Jr, and G. G. Moisen. 2007. Habitat classification modeling with incomplete data: pushing the habitat envelope. Ecological Applications 17(6):1714-1726.