A DATA-DRIVEN METHOD FOR SELECTING OPTIMAL MODELS BASED ON GRAPHICAL VISUALISATION OF DIFFERENCES IN SEQUENTIALLY FITTED ROC MODEL PARAMETERS

K S Mwitondi*[*1], R E Moustafa[2], and A S Hadi[3]

*[*1]Sheffield Hallam University, Faculty of Arts, Computing, Engineering and Sciences, Sheffield S1 1WB, UK*
Email: k.mwitondi@shu.ac.uk, *mwitondi@yahoo.com*
[2]George Washington University, Statistics Department, 2140 Pennsylvania Ave., NW, Washington DC, 20052, USA
Email: Shalash@gwu.edu, moustafa@dmining-technology.com
[3]The American University in Cairo, Egypt/Cornell University, 291 Ives Hall, Cornell University, Ithaca, NY 14853-3901, USA
Email: ahadi@aucegypt.edu, ali-hadi@cornell.edu

ABSTRACT

Differences in modelling techniques and model performance assessments typically impinge on the quality of knowledge extraction from data. We propose an algorithm for determining optimal patterns in data by separately training and testing three decision tree models in the Pima Indians Diabetes and the Bupa Liver Disorders datasets. Model performance is assessed using ROC curves and the Youden Index. Moving differences between sequential fitted parameters are then extracted, and their respective probability density estimations are used to track their variability using an iterative graphical data visualisation technique developed for this purpose. Our results show that the proposed strategy separates the groups more robustly than the plain ROC/Youden approach, eliminates obscurity, and minimizes over-fitting. Further, the algorithm can easily be understood by non-specialists and demonstrates multi-disciplinary compliance.

Keywords: Bayesian error, Data mining, Data visualisation, Decision trees, Domain partitioning, Optimal bandwidth, ROC curves, Visual analytics, Youden Index

1 INTRODUCTION

Choosing from a range of competing models is a common practice in predictive modelling in which the selection of the optimal model and its performance depends on a combination of factors. In classification, for instance, the consequences of misclassification largely depend on the true class of the object being classified and the way it is ultimately labelled. Generally, the performance of both parametric and non-parametric models depends exclusively on the chosen model, the sampled data, and the available knowledge for the underlying problem. For instance, the accuracy and reliability of, say, a medical test will depend not only on the diagnostic tools but also on the definition of the state of the condition being tested. Such variations make model complexity a natural challenge to data modelling (Mwitondi, 2010). Thus, when data sources, repositories, and modelling tools are shared, it is imperative to work out a unifying environment with the potential to yield consistent results across applications. Achieving this goal requires striking a balance between model accuracy and reliability across applications (Mwitondi & Said, 2011). This paper examines how multiple model performances can be used to devise a generalised strategy for attaining the foregoing balance in predictive modelling. Using a generic two-class scenario, it addresses the underlying issues relating to prediction errors and combines model generated numerals and graphics to decipher data patterns for optimality. More specifically, the paper sets off from conventional approaches for group separation, ROC curves (Egan, 1975) and the Youden Index (Youden, 1950), to propose an iterative algorithm for detecting separation levels based on estimated data densities. The paper is organised as follows. Section 2 provides an overview of the methods and simulations. Section 3 outlines the modelling strategy, implementation, results, and discussions. The concluding remarks and potential future applications are outlined in Section 4.

2 METHODS AND SIMULATIONS

The ultimate goals are to illustrate the nature of variation in the performance of various models given the random nature of the data used to train and test them and to devise a modelling strategy for optimising the model selection process. The illustrations and the strategy derive from the Bayesian rule as outlined in Berger (1985), the decision trees (DT) domain partitioning technique as described in Breiman et al. (1984), and the receiver operating characteristics (ROC) analysis as outlined in Egan (1975).

2.1 Allocation rule errors due to data randomness

As shown in Table 1, the total empirical error is typically associated with randomness due to the allocation region and randomness due to assessing the rule by random training and validation data (Mwitondi, 2003).

Table 1. Error types associated with domain-partitioning modelling (Source: Mwitondi, 2003)

POPULATION	TRAINING	CROSS VALIDATION	TEST
$\Psi_{D,POP}$	$\Psi_{D,TRN}$	$\Psi_{D,CVD}$	$\Psi_{D,TST}$

Thus, given that there are $X_{i=1,2,\ldots,N}$ data points in $Y_{k=1,2,\ldots K}$ different classes, the overall misclassification error is computed as the sum of the weighted probabilities of observing data belonging to a particular class given that we are not in that class. For instance,

$$\Psi_{D,TST} = \sum_{k=1}^{K} \sum_{i=1}^{N} P(C_k) P(X_i \in C_k | Y \notin C_k) \quad (1)$$

where C_k and $P(C_k)$ represent the partition region and the class priors respectively in a typical Bayesian context. Various approaches for minimising this error have been proposed - see, for instance, Reilly and Patino-Leal (1981), Wan (1990), Freund and Schapire (1997), and Mwitondi et al. (2002). A commonly acceptable practice is to vary the allocation rule in order to address specific requirements of an application. We illustrate the scenario based on decision trees as in Breiman et al. (1984) and ROC curves (Egan, 1975).

2.2 Decision trees modelling

Growing a tree amounts to sequentially splitting the data into, typically, two super sets, A and B, based on a single predictor at a time. The observations in A and B lie on either side of the hyper-plane $x_j = m$ chosen in such a way that a given measure of impurity is minimised. The splitting continues until an adopted stopping criterion is reached. Selecting an optimal model is one of the major challenges data scientists face. Breiman et al. (1984) propose an automated cost-complexity measure described as follows. Let the complexity of any sub-tree $f \in F$ be defined by its number of terminal nodes, L_t. Then, if we define the cost-complexity parameter $0 \leq \alpha < \infty$, the cost-complexity measure can be defined as

$$R_\alpha(f^\alpha) = R(f) + \alpha L_f \quad (2)$$

Let f_t be any branch of the sub-tree $f^{(1)}$, and define $R(f_t) = \sum_{t^* \in L_{\alpha,f_t}} R(t^*)$ where L_{α,f_t} represents the set of all terminal nodes in f_t. They further show that given t any non-terminal node in $f^{(1)}$, the inequality $R(t) > R(f_t)$ holds. It can be shown that for any sub-tree f_t we can define a measure impurity as a function of α as

$$R_\alpha(f_t) = R(f_t) + \alpha L_{f_t} \quad (3)$$

Typically, growing a large tree yields high accuracy but risks over-fitting while growing a small tree does the opposite. The measure of impurity will typically return different estimates for different values of α directly impinging on accuracy and reliability. One way of assessing model performance is to use ROC curves.

2.3 ROC curves analysis, optimality, and Youden Indexing

Without loss of generality, consider a binary medical diagnostic test scenario in which patients are tested for a particular disease and there are four possible outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In this case, the ROC curve is constructed based on the proportions

$$SST = \frac{N_{TP}}{N_{TP}+N_{FN}} \qquad \text{and} \qquad SPT = \frac{N_{TN}}{N_{TN}+N_{FP}} \tag{4}$$

where *SST* and *SPT* denote the sensitivity and specificity respectively and $\mathrm{SST} = 1 - \mathrm{SPT}$. N_{TP} and N_{FN} denote the number of those with the disease and who are diagnosed with it and those having the disease but cleared by the test respectively. Similarly, N_{TN} and N_{FP} are the number of those without the disease who test negative and those testing positive without having the disease, respectively. As with type I and II errors, the usefulness of a test cannot be determined by SST/SPT alone, and so a ROC analysis trade-off is needed. Assuming four possible outcomes, the ROC accuracy (ACCR) and error (ERR) are defined as

$$ACCR = \frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}} \leftrightarrow 1 - ACCR = ERR \tag{5}$$

If we denote the data by X and the set of class labels as $C_i = \{Y_1, Y_2\}$, the probability of accuracy can be computed as follows, where the integral is over both classes.

$$P(ACCR) = P(Y_i \in C_i) = \sum_{i=1}^{2} P(Y_i) \int P(X|Y_i)\, dx \leftrightarrow 1 - P(Y_i \in C_i) = P(ERR) \tag{6}$$

The main goal of predictive modelling is to maximise $P(ACCR)$ or equivalently minimise $P(ERR)$ consistently across applications. By appropriately costing each of the class allocation measures, we can make the outcome not only depend on the diagnostic tools and techniques but also on the definition of the state of the tested condition. For instance, one would rather set "low specificity" for a cancer diagnostic test, i.e., let it trigger on low-risk symptoms, than miss the symptoms. Our model implementations are focused on striking this balance. One way of determining the optimal cut-off point for the ROC curves is to use the Youden Index (Youden, 1950). Its main idea is that for any binary classification model with corresponding cumulative distribution functions $\mathrm{F}(*)$ and $\mathrm{G}(*)$, say, then for any threshold *t,* the relationship $\mathrm{SST}(t) = 1 - \mathrm{F(t)} \leftrightarrow \mathrm{SPT(t)} = \mathrm{G(t)}$ holds. We can then compute the index γ as the maximum difference between the two as

$$\gamma = max_t\{SST(t) + SPT(t) - 1\} = max_t\{G(t) - F(t)\} \tag{7}$$

Within a model, the Youden Index is the maximum differences between the true and false positive values and between competing models ordering of the indices highlights performance order.

3 MODELLING STRATEGY, IMPLEMENTATION, RESULTS, AND DISCUSSIONS

The modelling strategy is based on the methods described in Section 2 and seeks to facilitate the selection of optimal models based on consistency of performance. Two datasets, the Pima Indians diabetes data, 768 observations on 9 variables (NIDDK, 1990) and the Bupa liver disorders data, 345 observations on 7 variables (Forsyth, 1990), as described in Table2 are used. The former relate to females of at least 21 years old while the latter relate to blood tests for liver disorders sensitivity to excessive alcohol consumption. Because variations in α affect reliability, we trained and tested three decision tree models on each of the datasets.

Table 2. Data attributes for the Pima Indians diabetes and the Bupa liver disorders datasets

PIMA INDIANS DIBETES DATA		BUPA LIVER DISORDERS DATA	
NTP	Number of times pregnant	**MCV**	Mean corpuscular volume
PGC	Plasma glucose concentration	**ALKPHOS**	Alkaline phosphotase
DBP	Diastolic blood pressure	**SGPT**	Alamine aminotransferase
TSF	Triceps skin fold thickness	**SGOT**	Aspartate aminotransferase
BMI	Body mass index	**GAMMAGT**	Gamma-glutamyl transpeptidase
DPF	Diabetes pedigree function	**DRINKS**	Daily half-pint drink equivalents
AGE	Age (years)	**CLASS**	Binary target
CLASS	Binary Target		

3.1 Implementation and results

Graphical results for both datasets are presented in Figure 1. Left to right, the first two panels correspond to Pima ROC and predictive patterns while the remaining two correspond to those of the Bupa dataset.

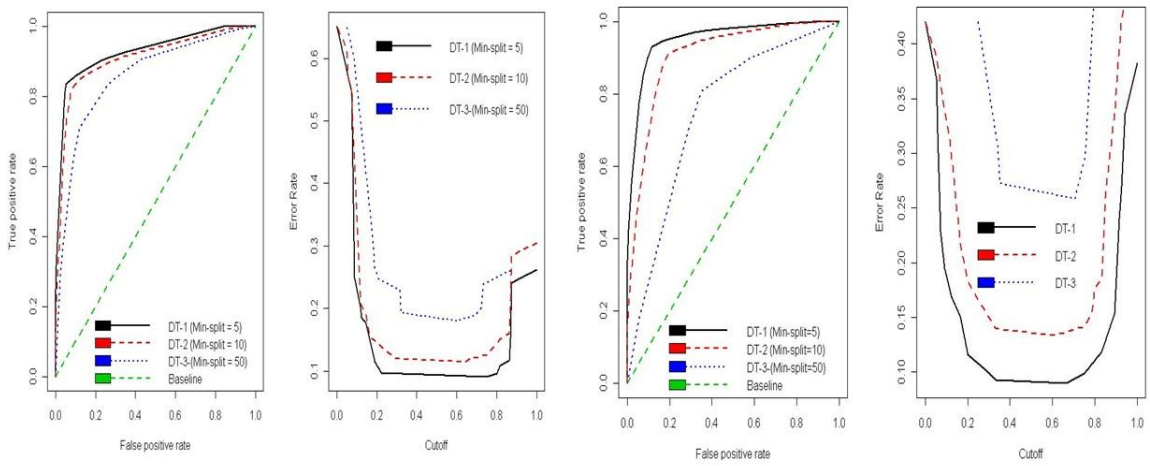

Figure 1. Pima (left) and Bupa(right) ROC curves and model over/fitting points

Based on the ROC convention, a classifier is optimal only if it yields results in the top left corner, given the set conditions. Thus, the performance ranking in both cases was DT-1, DT-2, and DT-3. The maximum differences (Youden Indices) between the TPR and FPR for each model, the areas under the curve, and the minimum prediction errors are highlighted in Table 3, agreeing with this ranking. Note that Bupa's gaps between DT1 and DT2 and between DT2 and DT3 are much wider than Pima's, implying that DT1 and DT2 performance on the Pima Indians data is almost indistinguishable. Further, repeated simulations are expected to vary depending on factors such as data sources and the settings defined in Section 2.2.

Table 3. Performance table for each of the three models on each of the two datasets

	PIMA-1	PIMA-2	PIMA-3	BUPA-1	BUPA-2	BUPA-3
YOUDEN INDEX	0.7838	0.7392	0.5907	0.8128	0.7169	0.4583
AREA UNDER THE CURVE	0.9273	0.9086	0.8556	0.9527	0.9069	0.7487
MINIMUM ERROR	0.0911	0.1146	0.1797	0.0899	0.1333	0.2580

Accuracy of the test is a function of how well it separates the groups, and it is assessed by the corresponding area under the curve. Traditionally, a 90%-100% area under the curve is considered excellent while anything about 60% and less is a failure. Because ROC curves may mask or over-fit data estimated parameters, we propose a strategy for enhancing the formulation of a generalised error.

3.2 Proposed strategy for optimal model selection

Typically, one classifier is preferred to another if it yields a higher class posterior probability than the other (Web, 2005; Mwitondi, 2003). We have shown that the rate at which the models out-perform each other is data-dependent, and so the fitting patterns may provide a good starting point in the search for optimality. If we assume a Bayesian error from a notional population on which the performance of a predictive model is assessed to be $\Psi_{B,POP} = \Psi_{D,TST}$ (data-dependent error), then the relationship below holds.

$$P\left(\Psi_{D,POP} \geq \Psi_{B,POP}\right) = 1 \leftrightarrow E\left[\Psi_{D,POP}\right] - \Psi_{B,POP} = E[\Delta] \geq 0 \tag{8}$$

We can then measure model reliability by tracking the quantity $Var[\Delta]$ as an indicator of stability/variability across models. The algorithm, described in the following simple steps, seeks to minimise the risk of over-fitting or under-fitting the data across applications.

Given a set of competing classifiers $\{C_j\}\ j = 1,2, \ldots K$
Extract the vectors $TP = X_i^T$ *and* $FP = X_i^F$
Set $D_i = X_i^T - X_i^F$
For j:=1:K
For $i := 1: i-1$
$DIFFS = D_{i+1} - D_i$
$TP_d = X_{i+1}^T - X_i^T$
$FP_d = X_{i+1}^F - X_i^F$
End For
Store the Differences DIFFS, TP and FP
End For
Set a long bandwidth vector (typically Gaussian) $\beta \in (1,0)$
While NOT END of β *Do*
Compute and plot the densities of D_i, TP_d *and* FP_d
End While
Examine the resulting plots and choose the one that best separates the groups
End.

Graphical illustrations of the algorithm results based on both datasets at different bandwidths are shown in Figure 2. The spiky lines at the foot of each of the four density panels represent a slightly noised univariate vector of TP and FP from left to right. Because the main purpose is to separate each of the two classes, we are interested not only in the positioning of the ROC curves as in Figure 1 but also in the sequential differences in the fitted parameters, which suggest that DT1 is suitable for the BUPA and DT2 for the PIMA data.

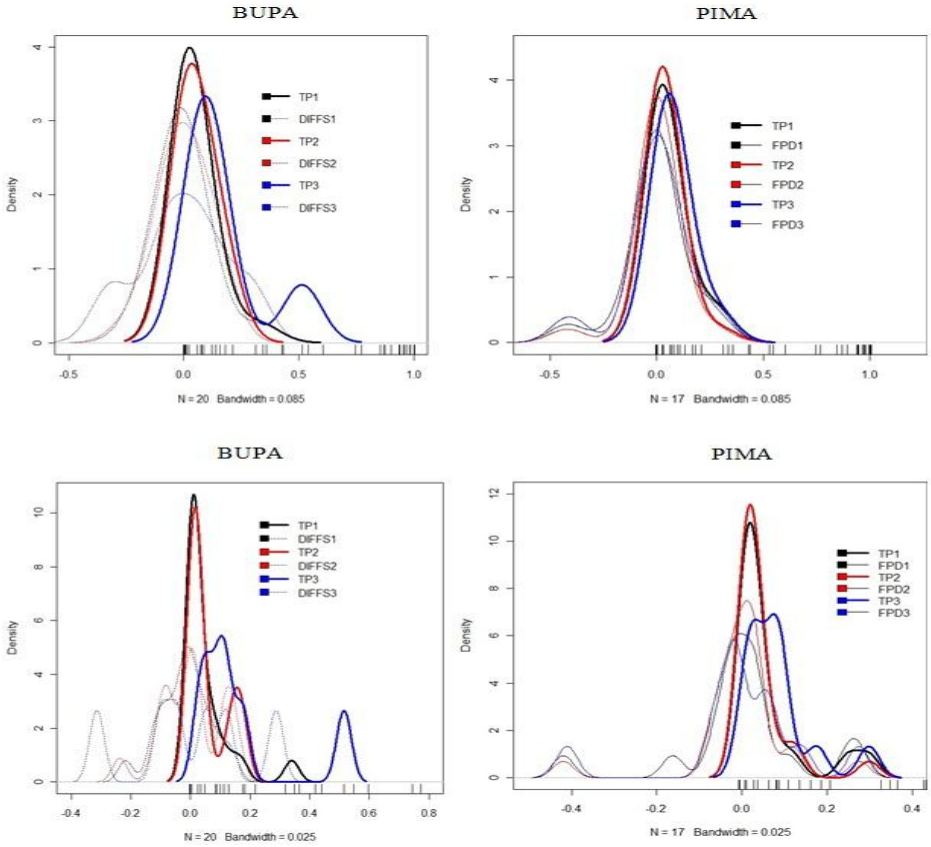

Figure 2. Differences between sequential TP and FP values and those of differences between them

The Gaussian kernel was used in approximating the differences in the algorithm. The optimal choice of the bandwidth is estimated as $b = \left(4\hat{\sigma}^5/3n\right)^{1/5} \approx 1.06\hat{\sigma}n^{-1/5}$ (Silverman, 1984) where sigma is the standard deviation of the samples n. Runs based on these optimal bandwidth values yielded similar results to those in Figure 2 at $b \leq 0.1$ suggesting DT1 for BUPA and DT2 for PIMA. Because each of the ROC curves in Figure 1 measures the probability that the corresponding model will rank a randomly chosen positive instance higher than it will rank a randomly chosen negative case, the patterns in Figure 2 provide an insight into the level of class separation and can be used to guide model selection.

4 CONCLUDING REMARKS AND POTENTIAL FUTURE DIRECTIONS

Selecting the "best" model from a potentially set of competing models is a conventional challenge in data science, and this paper sought to demonstrate an optimisation procedure for making that decision. Guided by the Bayesian rule, ROC curves, and the Youden Index, we empirically demonstrated the variation of the allocation rule using three decision tree models. For the purpose of addressing specific application requirements, a practical reality in a data sharing environment, we introduced a novel strategy for model selection. Based on graphical visualisation, the strategy seeks to help minimise data over-fitting and performance obscurity while remaining easily understood by non-specialists. The results from this paper serve to highlight the importance of paying attention to the allocation rules in Table 1 and the associated generalising error. The strategy can be adapted to other data-dependent domain-partitioning models, such as neural networks and support vector machines. The quantity $\mathrm{Var}[\Delta]$ can generally be accepted as a measure of performance that, for domain-partitioning purposes, we can align alongside similar measures, such as the ROC curves. The proposed strategy can readily be adapted to all applications of a binary nature or those that can be converted into such, and our results highlight novel paths towards tackling various real-life challenges in areas such as remote sensing, seismology, oceanography, ionosphere, and many others. Since error costing differs across applications, the decisions relating to model selection will typically remain application-specific. However, the proposed strategy provides prospects of interactivity and multi-disciplinary compliancy in a general data sharing framework

irrespective of the nature of the applications. We hope that this study will supplement previous studies that have focused on methods for selecting optimal models in our increasingly expanding cross-disciplinary research environment.

5 REFERENCES

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag.

Breiman, L., Friedman, J. Stone, C. J., & Olshen, R. A. (1984) *Classification and Regression Tree*, Chapman and Hall, ISBN-13: 978-0412048418.

Egan, J. P. (1975) *Signal Detection Theory and Roc Analysis,* Academic Press; ISBN-13 978-0122328503.

Freund, Y. & Schapire, R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55* (1), pp 119–139.

Forsyth, R. S. (1990) *PC/BEAGLE User's Guide*, BUPA Medical Research Ltd.

Mwitondi, K. S. (2003) *Robust Methods in Data Mining*, PhD Thesis, School of Mathematics, University of Leeds, Leeds, University Press.

Mwitondi, K. S. & Said, R. A. (2011) A step-wise method for labelling continuous data with a focus on striking a balance between predictive accuracy and model reliability. *International Conference on the Challenges in Statistics and Operations Research (CSOR).*

NIDDK (1990) *Pima Indians Diabetes Data*, National Institute of Diabetes and Digestive and Kidney Diseases.

Reilly, P. M. & Patino-Leal, H. (1981) A Bayesian Study of the Error-in-Variables Model. *Technometrics 23*, (3).

Silverman, B. W. (1986) Density Estimation for Statistics and Data Analysis. *Monographs on Statistics & Applied Probability*, Chapman and Hall, ISBN-13: 978-0412246203.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005) ROCR: Visualizing Classifier Performance in R. *Bioinformatics 21* (20), pp. 3940-3941.

Youden, W.J. (1950) Index for rating diagnostic tests. *Cancer 3*, pp 32-35.

(Article history: Available online 2 May 2013)