Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.12785/jsap/020312

A Data-based Method for Harmonising Heterogeneous Data Modelling Techniques Across Data Mining Applications

Kassim S. Mwitondi[1] and Raed A. T. Said[2]

[1]Sheffield HallamUniversity, Department of Computing, SheffieldS1 1WB,UK
[2]Al AinUniversity of Science and Technology, Al Ain, United Arab Emirates
*Email:*k.mwitondi@shu.ac.uk[1] , mwitondi@yahoo.com[1],*staff.raed.s@alainuniversity.ac.ae*[2],*raedsaeed@yahoo.com*[2]

Received: 23Apr. 2012, Revised: 14Nov. 2012, Accepted: 23 Feb. 2013
Published online: 1 Nov. 2013

Abstract: We propose an iterative graphical data visualisation algorithm for optimal model selection. The algorithm is implemented on three domain-partitioning techniques - decision trees, neural networks and support vector machines. Each model is trained and tested on the Pima Indians and Bupa Liver Disorders datasets with the performance being assessed in a multi-step process. Firstly, the conventional ROC curves and the Youden Indexare applied to determine the optimal model then sequential moving differences involving the fitted parameters - true and false positives – are extracted and their respective probability density estimations are used to track their variability using the proposed algorithm. The algorithm allows the use of data-dependent density bandwidths as tuning parameters in determining class separation across applications. Our results suggest that this novel approach yields robust predictions and minimizes data obscurity and over-fitting. The algorithm's simple mechanics which derive from the standard confusion matrix and built-ingraphical data visualisationand adaptive bandwidth featuresmake it multidisciplinary compliant and easily comprehensible to non-specialists. The paper's main outcomes are two-fold. Firstly, it combines the power of domain partitioning techniques on Bayesian foundations with graphical data visualisation to provide a dynamic, discernible and comprehensible information representation. Secondly, it demonstrates that by converting mathematical formulation into visual objects, multi-disciplinary teams can jointly enhance the knowledge of concepts and positively contribute towards global consistency in the data-based characterisation of various phenomena across disciplines.

Keywords: Bayesian Error, Data Mining, Decision Trees, Domain Partitioning, Data Visualisation, Neural Networks, Optimal Bandwidth, ROC Curves, Support Vector Machines, Youden Index

1 Introduction

Enhancements in data acquisition, manipulation and transmission are constantly increasing the appetite for data, information and knowledge consumption across applications. However, these developments also entail sophistications in capturing of the underlying rules upon which the data attributes interact as well as in the way modelling results are interpreted and shared across applications. In classification, for instance, the accuracy and reliability of, say, a medical testwill depend not only on the diagnostic tools but also on the definition of the state of the condition being tested. Typical applications employ unsupervised and supervised modelling techniques to detect meaningful data patterns based on information in data attributes – a process, generally described as data mining or knowledge discovery from data (KDD). Applications of data mining algorithms such as Neural Networks, Decision Trees, Support Vector Machines and other adaptive methods are well documented – see, for example, Kirkos et al., (200) and Sangita (2011). Applications of the methods support multi-disciplinary research in that the models can readily be applied across fields. For instance, the classification of wind speed in Sangita (2011) can be useful

not only in the field of power and energy, but also in environmental studies such as tracking the movement of particulates and other air pollutants in a specified region. The overall performance of a typical data modelling technique depends on the chosen model, the sampled data and the available knowledge for the underlying problem domain. In other words, knowledge extraction from data relies on a combination of factors which together determine the selection of the optimal model and its performance.

With most applications relying on disparate data sources, repositories and modelling tools it is quite imperative to try and work out a unifying environment with the potential to yield consistent results across applications(Mwitondi and Said, 2011). This paper applies disparate data modelling techniques to setsreal and simulated binary target datasets to highlight the inherent modelling complexities. It then goes on to use the conventional results as inputs into a novel application independent strategy for model optimisation and practically demonstrate the algorithm's multi-disciplinary compliance.

The paper's main objective is to contribute to the development of a data modelling environment in which information extracted from data is communicated to users via rigorous, unambiguous and universally comprehensible representation methods. Its main outcomes are two-fold. Firstly, it provides a discernible and comprehensible information representation through dynamic visualisation of modelling results - helping to enhance knowledge of concepts and parameters. Secondly, it shows that by converting mathematical formulation into visual objects, multi-disciplinary teams can jointly contribute towards global consistency in the data-based characterisation of phenomena in real-life applications. The paper is organised as follows. Section 0provides a general overview of the Bayesian rule; outlines the fundamental mechanics of selected predictive modellingtechniques and highlights their suitability to harmonisation.

2 Methods

The methods derive from the following key elements of data mining - as data (quite heterogeneous in nature), tools and techniques (with disparate mechanics and functionalities) and analytical and interpretational skills (quite often multi-disciplinary in nature). To illustrate the different forms of randomness in model training and testing, we use two datasets – the Pima Indians Diabetes and the Bupa Liver Disorders data. The former consists of 768 observations on 9 variables (NIDDK, 1990) relating to females of at least 21 years old - and the latter consists of 345 observations on 7 variables (Forsyth, 1990) - the first 5 relating to blood tests for liver sensitivity to excessive alcohol consumption. The ultimate goals are to illustrate the nature of randomness in predictive accuracy and reliability and propose a novel strategy for striking a balance between the two. The modelling strategy derives from the Bayesian rule as in Berger (1985) and the modelling techniques are adapted from Valiant (1984), Breiman*et al.*, (1984) and Egan (1975).

2.1 Bayesian foundations of allocation rules in predictive modelling

Analogical to the type I and type II errors in hypothesis testing, predictive modelling is associated with a loss function or a measure of the discrepancy between the true and estimated probabilities. Assuming that a correct classification incurs no loss, a prediction rule can be defined as $\frac{P(X|C_1)}{P(X|C_2)} > \frac{c_{21}P(C_2)}{c_{12}P(C_1)}$where $P(C_{1/2})$ are the class priors and $c_{21/12}$represent the cost of incorrectly allocating an observation to a class which also implies that $P(C|X) > \frac{c_{21}}{c_{12}+c_{21}}$. Webb

(2005) shows that the Bayesian decision rule for minimum risk or the total loss is the weighted sum in Equation 1where $\omega_{1/2}$ are the probabilities of misclassifying $C_{1/2}$ respectively.

$$\Psi = c_{12}P(C_1)\omega_1 + c_{21}P(C_2)\omega_2 \quad (1)$$

If the parameters in Equation 1 were known, computation of the loss function would be trivial. However, in most applications these parameters are estimated from data and as such they inevitably affect Ψ in different ways. As shown in Table 1, this is because the total empirical erroris affected by 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 DP modelling (Source: Mwitondi, 2003)

ALLOCATION RULE ERRORS DUE TO DATA RANDOMNESS			
POPULATION	TRAINING	CROSS VALIDATION	TEST
$\Psi_{D,POP}$	$\Psi_{D,TRN}$	$\Psi_{D,CVD}$	$\Psi_{D,TST}$

Whether we know the parameters or we estimate them from data the Bayesian logic stipulates that we update currently available (prior) information into new (posterior) information. Thus, given that there are $X_{i=1,2,\dots,N}$ data points in $Y_{k=1,2,\dots K}$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. That is,

$$\Psi_{D,TST} = \sum_{k=1}^{K}\sum_{i=1}^{N} P(C_k)P(X_i \in C_k | Y \notin C_k) \quad (2)$$

where$C_k$and $P(C_k)$represent the partition region and the class priors respectively.Minimising the Bayesian error has always been central to predictive modelling – see, for instance, Reilly and Patino-Leal(1981), Wan (1990), Freund and Schapire (1997) and Mwitondi*et al.* (2002). If we let the Bayesian error from a notional population in Table 1 on which the performance of the model is assessed to be $\Psi_{B,POP} = \Psi_{D,TST}$, then

$$P\left(\Psi_{D,POP} \geq \Psi_{B,POP}\right) = 1 \leftrightarrow E\left[\Psi_{D,POP}\right] - \Psi_{B,POP} = E[\Delta] \geq 0 \quad (3)$$

Our focus can then be on minimising the unknown quantity$\Delta$and measuring and tracking its reliability across applications – a generic task applying to all data-dependent domain-partitioning models. In most applications a commonly acceptable practice is to vary the allocation rule in order to address specific requirements of an application. The applications in this paper derive from this convention. In the next exposition we examine some of the fundamental mechanics of the paper's adopted modelling methods.

2.2 Predictive modelling and performance assessment

Class allocation and associated loss are two of the key elements of predictive modelling. We illustrate them via Decision Trees, Neural Networks, Support Vector Machines.

2.2.1 Decision Trees (DT)

Given training data (x_i, y_i); $x_{i=1,2,\dots,N} \in \mathbb{R}^p$ and $y_{i=1,2,\dots,N} \in \{A, B\}$, growing a tree amounts to sequentially splitting the data into the two subsets A and B based on, typically, a single predictor at a time. If we denote the observations at any arbitrary tree node by N^* and the number of cases at the node by $\eta \in N^*$ then given the choice of a variable, say, $x_i \in X$ and the threshold on it, say, *m*, split the data into A and B such that

$$\begin{cases} A = \{\eta \in N^* : x_i \leq m \\ B = \{\eta \in N^* : x_i > m \end{cases} \quad (4)$$

The observations in A and B lie on either side of the hyper-plane $x_i = m$ chosen in such a way that a given measure of impurity is minimised. Without attempting to optimise the whole tree, this 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

$$R_\alpha(f^\alpha) = R(f) + \alpha L_f \quad (5)$$

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 of impurity as a function of α as

$$R_\alpha(f_t) = R(f_t) + \alpha L_{f_t} \quad (6)$$

Our description of the mechanics of decision trees modelling seeks to highlight the main issues which data scientists must be aware of. In particular, growing a large tree will yield high accuracy but risks over-fitting the data while growing a small tree yields low accuracy and may under-fit the data. The measure of impurity in Equation 6 returns different estimates for different values of α directly impinging on accuracy and reliability.

2.2.2 Neural Networks (NN)

A neural networks model can generally be viewed as a multi-stage predictive system that adapts its structurein accordance with the flow of data inputs and their associated weights through the structure (Ripley, 1996). Typically, the model consists of an input and output layer with one or more hidden layers between them. Mathematically, an NN model consists of a sequence of nested functions each being defined by the preceding one via a weighted sum and activation function. We adopt a simple description of neural networks as a sophisticated version of non-linear statistical models (Hastie *et al.*, 2001) and illustrate the mechanics of NN from a binary target classification perspective as follows. Given a training dataset (x_i, y_i); $x_{i=1,2,\ldots,N} \in \mathbb{R}^p$ and $y_{i=1,2,\ldots,N} \in \{0,1\}$, we can initialise a set of weights $w_i \in \mathbb{R}^p$and iteratively keep updating them in search of a set that fits the training data well. Theseconventional mechanics of NN – also known as the perceptron learning rule (Ripley, 1996) - entail adapting the weights to the differences between the desired and actual model target output. Cross-entropy or deviance is a commonly used measure of impurity for determining the error and, for the binary case, it can be defined as

$$\mathcal{D} = -\sum_{i=1}^{N} [y_{i1} \log[f_1(x_i)] + y_{i2} \log[f_2(x_i)]] \quad (7)$$

with the corresponding NN classifier being $\mathcal{C}(x) = \text{argmax}_{k=1,2} f_{k=1,2}(x)$. Typically, NN minimise the measure via back-propagation which entails propagating prediction errors backwards from the output nodes into the input nodes. That is, the network calculates its error gradient based on the updated weights: $\Delta w_i = \nu * [E(y_i) - \hat{y}_i] x_i$ where ν is the learning rate and $E(y_i)\text{and}\hat{y}_i$ are the expected and actual target outcomes respectively. NN modelling is associated

with a number of issues – mainly the initial set of weights and data over-fitting which is the subject of subsequent sections in this paper.

2.2.3 Support Vector Machines (SVM)

Support Vector Machines (Vapnik, 1995; Cortes and Vapnik, 1995) represent another set of domain-partitioning methods which relies on training information to allocate new cases to known groups. Its basic idea is to map data examples as points in multi-space dimension separated by a hyper-plane or set of hyper-planes such that gap of disparity between distinct groups is as wide as possible. Group allocations are carried out by mapping all new cases onto the same space in accordance to which group they are predicted to belong. Generally, given a set of data (x_i, y_i); $x_{i=1,2,\dots,N} \in \mathbb{R}^p$ and $y_{i=1,2,\dots,N} \in \{-1,1\}$ define a hyper-plane $\{x : f(x) = x^T\beta + \alpha = 0\}$where β is a unit vector and $\|\beta\| = 1$ – implying that the distance from the origin to x is unit. The general allocation rule is therefore

$$R(x) = \text{sign}[x^T\beta + \alpha] \quad (8)$$

which gives the signed distance from x to the hyper-plane defined above. For separable classes we can find a hyper-plane $f(x) = x^T\beta + \alpha$ with $y_i f(x_i) > 0\ \forall_i$making it possible to find a hyper-plane with the widest gap between the two classes. Consequently, subject to $y_i(x_i^T\beta + \alpha) \geq G$,the corresponding optimisation problem is defined as

$$\max_{\beta,\alpha\|\beta\|=1} G \leftrightarrow \min_{\beta,\alpha}\|\beta\| \quad (9)$$

Note that $G = {}^1/_{\|\beta\|}$lies on either side of the plane – hence the gap is $2G = {}^2/_{\|\beta\|}$ wide. In the case of overlapping classes the above constraint can be modified either as

$$y_i(x_i^T\beta + \alpha) \geq G - \epsilon_i \text{ or } y_i(x_i^T\beta + \alpha) \geq G(1 - \epsilon_i) \quad (10)$$

wherethe slack variable $\epsilon_i \geq 0\ \forall_i$ and $\sum_{i=1}^{N} \epsilon_i \leq \tau$ (a constant).Typically, the two formulations in Equation 10 yield different solutions. The mechanics of SVM derive from $y_i(x_i^T\beta + \alpha) \geq G(1 - \epsilon_i)$ in which ϵ_i is the proportion of incorrect predictions (Hastie *et al.*, 2001). Since predictions errors arise when $\epsilon_i > 0$, bounding the sum of epsilon will bound $\Psi_{D,TRN}$. Hence, the choices of ϵ_i and τ are crucial to the overall performance of SVM and therefore the subsequent sections address performance-related issues.

2.3 Performance assessment using ROC curves analysis

The performance of all the foregoing techniques can be assessed by using ROC curves (Egan, 1975) which can be described using a simple binary medical diagnostic test scenario. If N 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) - then the ROC curve can be constructed based on the proportionsin Equation 11.

$$SST = \frac{N_{TP}}{N_{TP}+N_{FN}} \text{ and } \quad SPT = \frac{N_{TN}}{N_{TN}+N_{FP}} \quad (11)$$

where*SST*and *SPT*denote the sensitivity and specificity respectively and $\text{SST} = 1 - \text{SPT}$.$N_{TP}$ 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.

2.3.1 Maximising accuracy (minimising error)

ROC curves are used in selecting potential optimal (or discard non-optimal) models based only on the parameters ω_1 and ω_2 in Equation 1 and not on the cost and class distribution parameters. Krzanowski and Hand (2009) demonstrate various ways in which ROC curves can be used as performance measures by focusing on *inter-alia* statistical tests for ROC curves and their summary statistics.Assuming that a model will yield four possible outcomes – true positive, false positive, true negative and false negative the ROC accuracy (ACCR) and the corresponding error (ERR) are defined as in Equation 12.

$$ACCR = \frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}} \leftrightarrow 1 - ACCR = ERR \quad (12)$$

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

$$P(ACCR) = \sum_{i=1}^{2} P(Y_i) \int P(X|Y_i)\, dx \leftrightarrow 1 - P(Y_i \in C_i) = P(ERR) \quad (13)$$

The main goal of predictive modelling is to maximise $P(ACCR)$ or, equivalently, minimise$P(ERR)$consistently across applications which is basically model optimisation.

2.3.2 Area under the curve (AUC), model optimalityand Youden indices

The area under the ROC curve represents a measure of discriminatory performance with 1 corresponding to a perfect fit and 0.5 (below the baseline - no better than a random guess).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 $\mathrm{SST}(t) = 1 - \mathrm{F}(\mathrm{t}) \leftrightarrow \mathrm{SPT}(\mathrm{t}) = \mathrm{G}(\mathrm{t})$ holds. We can then compute the index γ as the maximum difference between the two functions as shown below.

$$\gamma = max_t\{SST(t) + SPT(t) - 1\} = max_t\{G(t) - F(t)\} \quad (14)$$

Within a model, the Youden index is the maximum differences between the true and false positives values and between competing models ordering of the indices highlights performance order. Thus, the index provides a handy tool in a shared data environment.

3 Implementation and proposed modelling strategy

To facilitate the selection of optimal models based on consistency of performance, we implement a two-stage analytical process. That is, we apply the three models and assess their performance using conventional methods followed by an implementation of the proposed model selection strategy based on the methods described in Section 0.

3.1 Conventional implementation

Implementation was based on three different decision tree sizes, three different architectures of feed-forward neural networks and three different support vector machines models. For the purpose of making an optimal choice from a set of competing models, we adopted simple distinguishing features in each set of models. We adapt the increasingly common approach to the representation of attributes - data visualisation as defined in Steele and Iliinsky (2010).

Implementation was in R2.1.3.0 (2011) and ROCR (Sing *et al.*, 2005) procedures and so the models' distinguishing features derive from the conventions embedded in those procedures. The adopted features were the minimum number of splits (DT), the number of units in the hidden layer (feed-forward NN)and the cost of constraints violation which represents the constant of the regularisation term in the Lagrange formulation (SVM). The ROC performances of the three sets are presented inFigure 1.

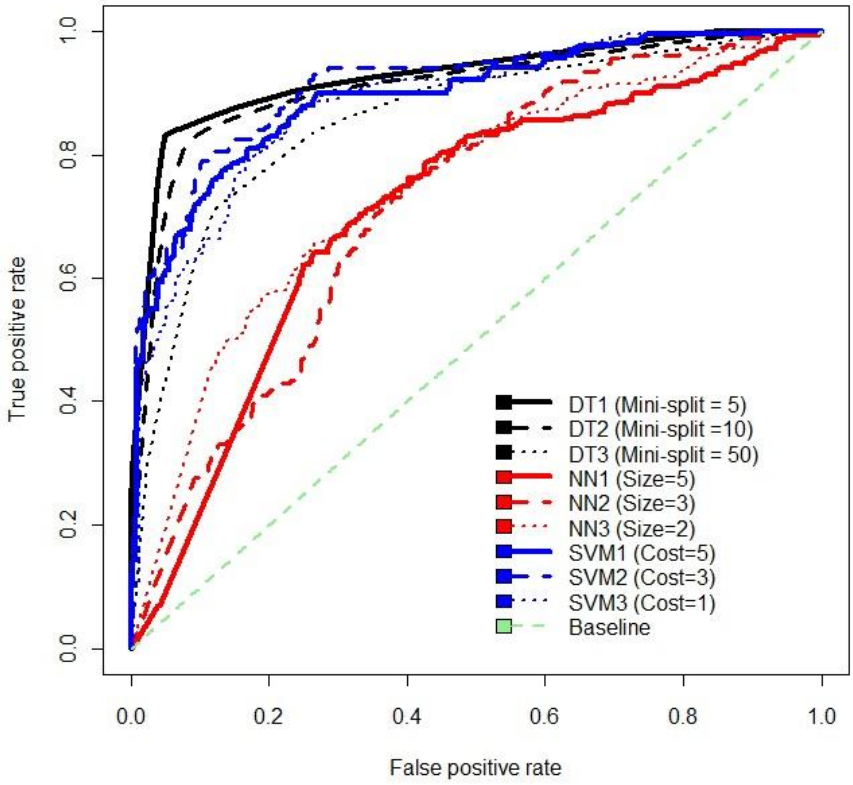


Figure 1: Combined DT, NN and SVM ROC curves for the Pima Indians data

The plots were generated using R2.1.3.0 (2011) code and ROCR (Sing *et al.*, 2005) embedded procedures. Each of the ROC curves measures the probability that the corresponding model will rank higher a randomly chosen positive instance than it will rank a randomly chosen negative case. The diagonal line is the baseline – the decision that could be taken without relying on the model which amounts to a random guess. In this case, the NN models were consistently out-performed by the other two. The Youden indices and areas under the curve (AUC) corresponding to the ROC curves in Figure 1 are given in Table 2. Based on AUC, the models are ranked DT1, SVM2, DT2, SVM3, SVM1, DT3, NN3, NN2 and NN1 while the Youden Index ($max_t\{G(t) - F(t)\}$) ranks them as DT1, DT2, DT3, SVM3, SVM1, NN3, SVM2, NN1 and NN2.

Table 2: Model performance parameters for the Pima Indians data

MODEL	MAXIMUM (TPR-FPR)	AREA UNDER THE CURVE
Decision Tree – 1 (DT1)	0.7838209	0.9272873
Decision Tree – 2 (DT2)	0.7391642	0.9086045
Decision Tree – 3 (DT3)	0.5906866	0.8555858

Neural Networks – 1 (NN1)	0.377791	0.7087612
Neural Networks – 2 (NN2)	0.3649254	0.7171828
Neural Networks – 3 (NN3)	0.3935224	0.7446828
Support Vector Machines – 1 (SVM1)	0.4082687	0.8937571
Support Vector Machines – 2 (SVM2)	0.3863284	0.9201000
Support Vector Machines – 3 (SVM3)	0.4255821	0.8955286

Models with exactly the same settings were trained and tested on the Bupa Liver Disorders data. Unlike in the previous case in which we had a much larger dataset with class priors $P(C_1 = 0.651)$and $P(C_2 = 0.349)$, the Bupa dataset was almost half the size of the Pima data but with slightly balanced priors at $P(C_1 = 0.579)$ and $P(C_2 = 0.421)$. The combined outputs from the three sets of models are graphically presented inFigure 2.

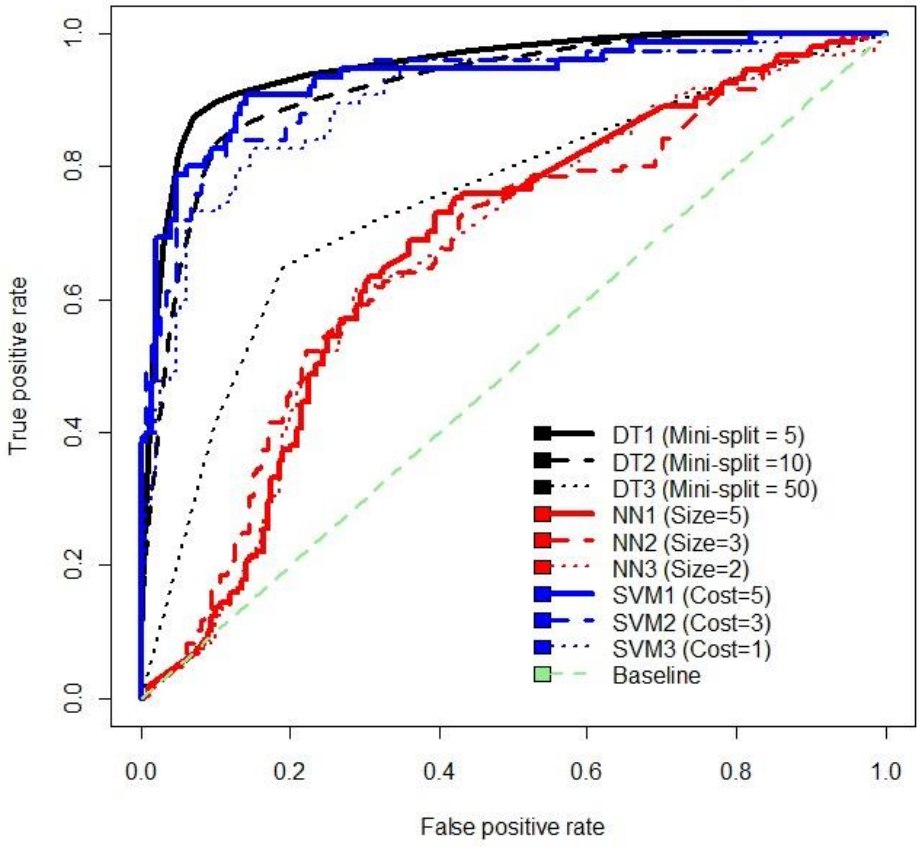


Figure 2: Combined DT, NN and SVM ROC curves for the Bupa Liver Disorders data

The model performance parameters - Youden indices and AUCs corresponding to the ROC curves in Figure 2 are given inTable 3. Based on AUC, the models are ranked in the order DT1, SVM1, DT2, SVM2, SVM3, NN1, DT3, NN3 and NN2while the Youden Index ranks them as DT1, DT2, SVM1, SVM2, SVM3, DT3, NN1, NN2 and NN3.

Table 3: Model performance parameters for the Bupa Liver Disorders data

MODEL	MAXIMUM (TPR-FPR)	AREA UNDER THE CURVE
Decision Tree – 1 (DT1)	0.7820896	0.9492241
Decision Tree – 2 (DT2)	0.7391642	0.9205172
Decision Tree – 3 (DT3)	0.5906866	0.7486724
Neural Networks – 1 (NN1)	0.4508621	0.7810517
Neural Networks – 2 (NN2)	0.3955172	0.7210345
Neural Networks – 3 (NN3)	0.3672414	0.7251034
Support Vector Machines – 1 (SVM1)	0.7266667	0.9290667
Support Vector Machines – 2 (SVM2)	0.7133333	0.9172444
Support Vector Machines – 3 (SVM3)	0.6600000	0.8863111

Due to the heterogeneous nature of the mechanics of each model, the rankings can only highlight the issue of training and testing models on inherently random data discussed earlier. Although it is imperative to assume that good training data will be used, it is always possible for this assumption to be violated due to factors such as variations in data sources and modelling practices which may lead to variations in drawn conclusions. In a multiple simulation binary example with balanced class priors, a 25% change in the standard deviation of the training data led to an average of 14% in the Youden indices and 10% in the ROC cut-off points. The two panels in Figure 3 exhibit comparisons between the NN and SVM models (LHS) and a DT within model comparison (RHS).

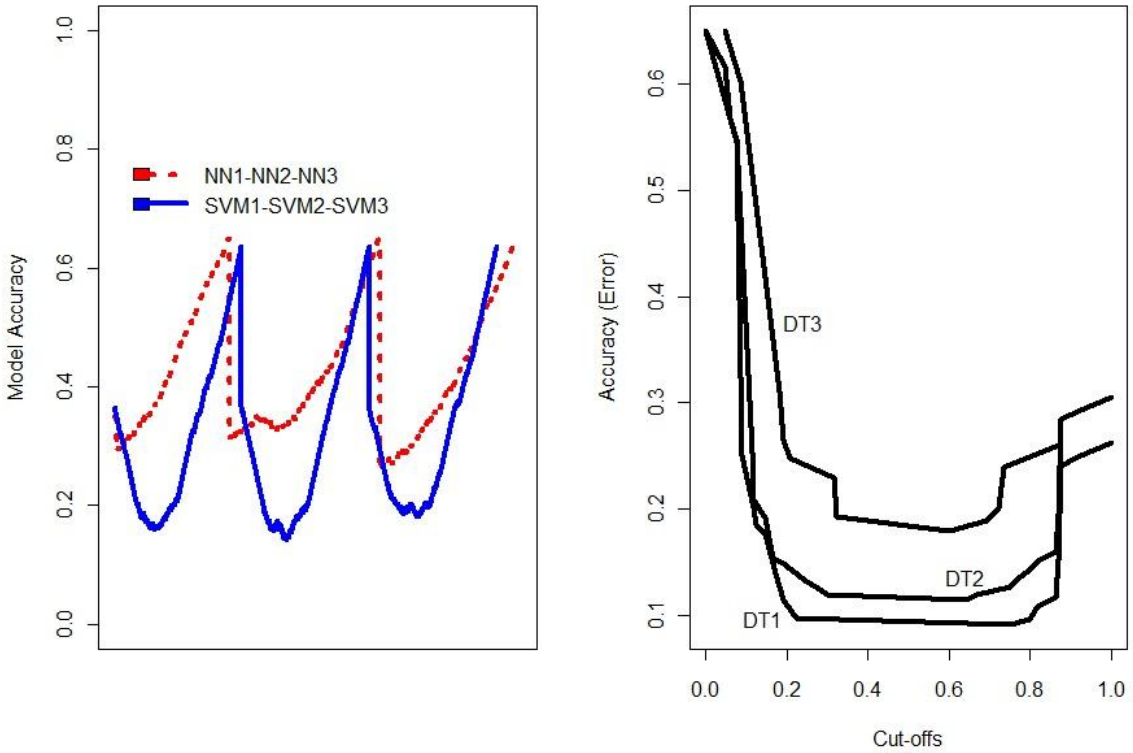


Figure 3: Error rates versus cut-off points for the Pima Indians data

Such variations make model complexity a natural challenge to data modelling (Mwitondi and Said, 2011) and impinge on the model accuracy (error) and reliability as shown inEquations 12,13 and3. Typically, repeated runs will vary depending on factors such as data sources and specific

model settings and so the over/fitting patterns may provide a good starting point in the search for optimality. In the following exposition, we propose a novel strategy aimed at enhancing the generic formulation in Equation 3.

4 Implementation of the proposed strategy

Our strategy for implementing the aforementioned domain-partitioning models and the procedure for selecting the optimal modelare based on simple mechanics. The main idea is to fit multiple models and select the model which yields consistent density plots from the step-wise differences in the ROC parameters. The algorithm's mechanics are as follows.

Given a set of competing classifiers $\{C_j\}\ j = 1, 2, \dots K$

Extract the vectors $TP = X_i^T,\ FP = X_i^F$

Set $D_i = X_i^T - X_i^F$

For j:=1:K

For $i := 1 : i - 1$

$$TP_d = \left(X_{i+1}^T - X_i^T\right) \Big/ \sigma_{N.T}$$

$$FP_d = \left(X_{i+1}^F - X_i^F\right) \Big/ \sigma_{N.F}$$

$$DIFFS = \left(D_{i+1} - D_i\right) \Big/ \sigma_{N.D}$$

$$AUC = AUC_{i+1}$$

End For

Store AUC and the differences DIFFS, TP and FP

End For

Set a long bandwidth vector (typically Gaussian) $\nu = \left[0 < \left(4\hat{\sigma}^5 / 3N\right)^{1/5} \leq 1\right]$

While NOT END of β ***Do***

Compute and plot the densities of DIFFS, TP_d ***and*** FP_d

End While

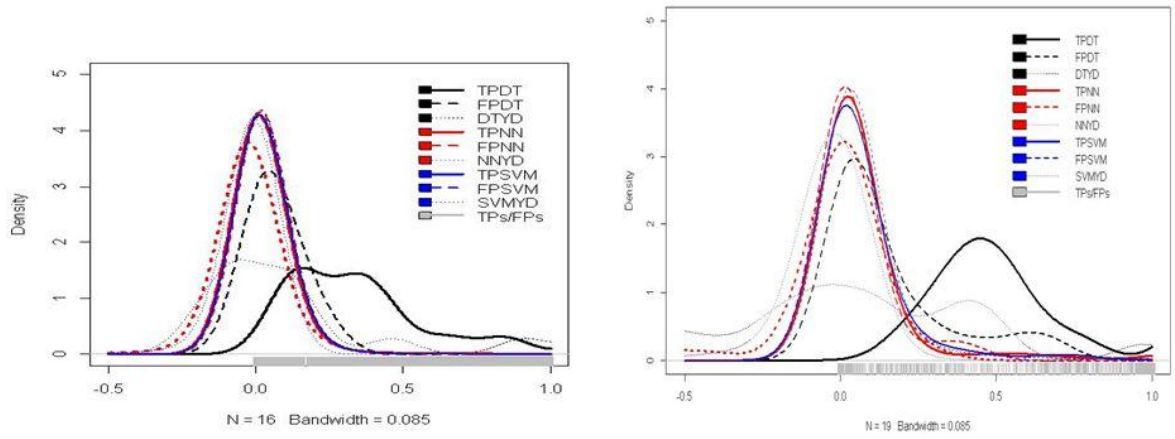
Examine the resulting plots and select C_j ***with the best group separation patterns.***

End.

The algorithm tracks the sequential differences between fitted TPs and FPs from which the Youden Index derives. The Gaussian kernel is used to approximate the differences in the algorithm with the optimal choice based on Silverman's optimal bandwidth criterion (Silverman, 1984) where σ is the standard deviation of the samples *N*. The foregoing algorithm adapts to specific applications via the adopted loss function.

The algorithm was tested on the best model from each of the three sets. For the Pima data those were DT1, NN1 and SVM2 while for the Bupa data it was DT1, NN1 and SVM1. At higher bandwidths the algorithms are indistinguishable as they reveal predominantly uni-modal features of the sequential differences. Graphical density patterns corresponding to each of the three models at bandwidth 0.085 are shown in Figure 4.TP and FP in the legend correspond to "true" and "false" positives respectively, DT, NN and SVM represent the models while YD highlights the

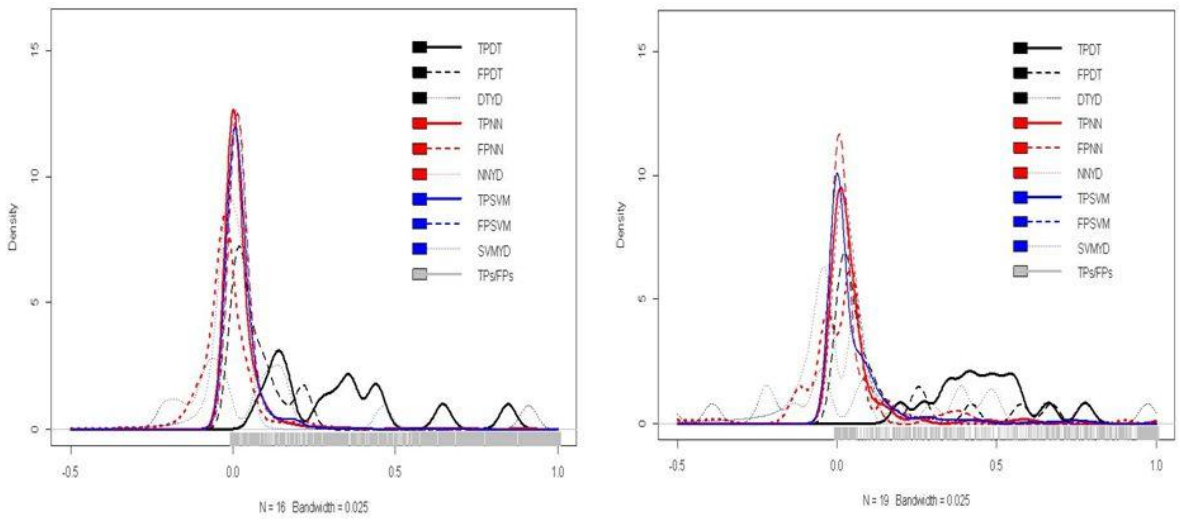
correspondence to the Youden Index. The greyed block at the foot of each plot (or TPs/FPs in the legend) represents a sorted vector of true and false positives as detected by all three models. These plots suggest that in both cases DT out-performs NN and SVM in capturing the natural data structures.



PIMA: Bandwidth = 0.085 **BUPA: Bandwidth = 0.085**

Figure 4: Density patterns for the Pima and Bupa sequential differences at bandwidth 0.085

As the bandwidth decreases the multi-modal features become more prominent and the performances of the three models become increasingly distinguishable. The graphical patterns in Figure 5detected at bandwidth 0.025 confirm the superiority of DT in both applications. These features are interesting because of their departure from the highly indistinguishable features between DT and SVM on Pima data with heavily unbalanced priors. Repeated runs of the models on simulated data confirmed this behaviour.



PIMA: Bandwidth = 0.025 **BUPA: Bandwidth = 0.025**

Figure 5: Density patterns for the Pima and Bupa sequential differences at bandwidth 0.025

The patterns in both panels of Figure 5provide an insight into the level of class separation and can be used to guide model selection. Further, ROC curves exhibit classifier performance irrespective of the parameters in Equation 1 which, in most applications, are unknown and must be estimated from the data. Thus, the algorithm is adaptable to all situations in which we can simultaneously train and test multiple models on available data. Note that we have confined applications to a two-class scenario. Extensions to multi-class scenarioscan be done in a number of ways. For example, for largely non-overlapping small number of classes, an indicator variable splitting the multi-classes into two supersets can be used. Alternatively, internal branching as in Zhu *et al.,* (2006) can be used.

5 Concluding remarks and potential future directions

This paper dwelled on the behaviour of fitted parameters from DT, NN, SVM and related performance assessment techniques - namely, ROC curves and the Youden Index. The rationale for proposed strategy derived from the context of the conceptual framework in Section 0. Typically, all three models yield a target output of the probability that the new observation is allocated to the correct class and so not only the probability target output must fulfill the conditions $0 \leq Y_k \leq 1$and $\sum_{k=0}^{K} Y_k = 1$. However, the magnitudes of these probabilities and other parameters vary with model complexity.

Our proposed algorithm sought to address the foregoing issues. Following implementation on the Pima and Bupa datasets, repeated simulations (at different bandwidths) revealed discernible class separations making it easy to interpret the results.It was shown that variability of delta can be used in conjunction with other measures of performance such as the ROC curves and we demonstrated the algorithm's capabilities in trackingit across models which helps minimise data obscurity and data over-fitting. While its mechanics rely on learningallocation rules from data, the rules generated will continue to require human intervention – preferably in the form of quantifiable human expert knowledge.

Over the last few decades, these developments have led to new data intensive multi-discipline applications in bioinformatics, business intelligence, data science etc. Yet, model optimisation will continue to be a major challenge among the data mining and data science communities. The strategy proposed in this paper can easily be extended to accommodate needs of a wide of applications. Since error costing differs across applications, the decisions relating to model selection remains application-specific. Thus, we see novel paths towards tackling new challenges in remote sensing, seismology, oceanography, ionosphere and many others. In space science, for instance, the relationships between the variations in the ionosphere and the climate can only be confirmed through rigorous and reliable predictions. We hope that this study will supplement previous studies on model selection methods and contribute towards cross-disciplinary research and help towards successful delivery of complex information to non-technical audiences.Extensions of this work also relies on the developments in data and modelling archives, which may provide scope for automating the selection criteria of C and the algorithm's tuning parameters.

References

[1] Breiman, L., Friedman, J. Stone, C. J. and Olshen, R. A. Classification and Regression Trees; Chapman and Hall, ISBN-13: 978-0412048418, (1984).

[2] Cortes and Vapnik, Support-vector networks; Machine Learning, Kluwer Academic Publishers, **20**, 273-297 (1995).

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

[4] Freund, Y. and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting; Journal of Computer and System Sciences, **55**, 119–139(1997).
[5] Forsyth, R. S., PC/BEAGLE User's Guide; BUPA Medical Research Ltd, (1990).
[6] Hastie, T., Tibshirani, R. and Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer, ISBN-13: 978-0387848570, (2001).
[7] Kirkos, E., Spathis, C. and Manolopoulos, Y., Support vector machines, Decision Trees and Neural Networks for auditor selection; Journal of Computational Methods in Science and Engineering, **8**, 213-224 (2008).
[8] Krzanowski, W. J. and Hand, D. J. , ROC Curves for Continuous Data; Chapman and Hall, (2009).
[9] Mwitondi, K. S. Robust Methods in Data Mining; PhD Thesis; School of Mathematics, University of Leeds; Leeds; University Press,(2003).
[10] Mwitondi, K. S and Ezepue, P. O. How to appropriately manage mathematical model parameters for accuracy and reliability: A case of monitoring levels of particulate emissions in ecological systems; International Conference on Mathematical Modelling of Some Global Challenging Problems in the 21st Century; Proceedings of NMC COMSATS Conference on Mathematical Modelling of Global Challenging Problems - 26th-30th, Nov. 2008, ISBN 978-8141-11-0, 24-36 (2008).
[11] Mwitondi, K. S. and Said, R. A., 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); 08th -10th March -2011, Kuwait City, (2011).
[12] Mwitondi, K. S., Taylor, C. C. and Kent, J. T., Using Boosting in Classification; Proceedings of the Leeds Annual Statistical Research (LASR) Conference; July 2002; pp. 125 – 128, Leeds University Press, (2002).
[13] NIDDK Pima Indians Diabetes Data; National Institute of Diabetes and Digestive and Kidney Diseases, (1990).
[14] R2.1.3.0, R Version 2.13.0 for Windows; R Foundation for Statistical Computing, (2011).
[15] Reilly, P. M. and Patino-Leal, H. A Bayesian Study of the Error-in-Variables Model; Technometrics, **23**, (1981).
[16] Ripley, B. Pattern recognition and neural networks; Cambridge University Press, ISBN-13: 978-0521717700, (1996).
[17] Sangita, B. P., Use of Support Vector Machine, decision tree and Naive Bayesian techniques for wind speed classification; In the Proceedings of the International Conference on Power and Energy Systems (ICPS), 22-24 Dec. 2011, Indian Institute of Technology, Madras, Chennai-36, INDIA, pp 1- 8, IEEE, ISBN 978-1-4577-1510-5, (2011).
[18] Silverman, B. W., Density Estimation for Statistics and Data Analysis; Chapman and Hall - Monographs on Statistics & Applied Probability; ISBN-13: 978-0412246203(1986).
[19] Sing, T., Sander, O., Beerenwinkel, N. andLengauer, T., ROCR: Visualizing Classifier Performance in R; Bioinformatics, **21**, 3940-3941 (2005).
[20] Steele, J. and Iliinsky, N. Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice); O'Reilly Media, ISBN-13: 978-1449379865(2010).
[21] Valiant, L. G. A theory of the learnable. Communications of the ACM 27, 1134–1142(1984).
[22] Vapnik, V., The Nature of Statistical Learning Theory. Springer-Verlag, ISBN 0-387-98780-0(1995).
[23] Wan, E.A., Neural Network Classification: A Bayesian Interpretation; Neural Networks, IEEE Transactions on Neural Networks, **1**, 303–305 (1990).
[24] Webb, A., Statistical Pattern Recognition; Wiley, ISBN 0-470-84514-7, (2005).
[25] Youden, W.J., Index for rating diagnostic tests. Cancer, **3**, 32-35 (1950).