Confident Interpretation of Bayesian Decision Tree Ensembles for Clinical Applications

Vitaly Schetinin, *Member, IEEE*, Jonathan E. Fieldsend, *Member, IEEE*, Derek Partridge, Timothy J. Coats, Wojtek J. Krzanowski, Richard M. Everson, Trevor C. Bailey, and Adolfo Hernandez

***Abstract*—Bayesian averaging (BA) over ensembles of decision models allows evaluation of the uncertainty of decisions that is of crucial importance for safety-critical applications such as medical diagnostics. The interpretability of the ensemble can also give useful information for experts responsible for making reliable decisions. For this reason, decision trees (DTs) are attractive decision models for experts. However, BA over such models makes an ensemble of DTs uninterpretable. In this paper, we present a new approach to probabilistic interpretation of Bayesian DT ensembles. This approach is based on the quantitative evaluation of uncertainty of the DTs, and allows experts to find a DT that provides a high predictive accuracy and confident outcomes. To make the BA over DTs feasible in our experiments, we use a Markov Chain Monte Carlo technique with a reversible jump extension. The results obtained from clinical data show that in terms of predictive accuracy, the proposed method outperforms the maximum *a posteriori* (MAP) method that has been suggested for interpretation of DT ensembles.**

***Index Terms*—Bayes procedures, Monte Carlo method, trees, uncertainty.**

I. Introduction

The assessment of uncertainty of decisions in safety-critical applications such as medical diagnostics etc., is of crucial importance. In general, uncertainty is a tradeoff between the amount of data available for training, the diversity of decision models, and their predictive accuracy [1]–[6]. The interpretability of classifiers can also give useful information to domain experts responsible for making reliable classifications. For this reason, decision trees (DTs) are attractive classification models for experts [2]–[7].

The main idea of using DT models is to recursively partition data points in an axis-parallel manner. Such models provide natural predictor selection and uncover the most important predictors for the classification. The resultant DT classification models are easily interpretable by users. By definition, DTs consist of splitting and terminal nodes, which are also known as tree leaves. DTs are said to be binary if the splitting nodes ask a specific question and then, divide the data points into two disjoint subsets, say the left or the right branch. The terminal node assigns all data points falling in that node to the majority class of the training data points that reach this terminal node. Within a Bayesian framework, the class posterior distribution is calculated for each terminal node [2]–[6]. An optimal outcome of decision models can be achieved by an averaging technique based on Bayesian Markov Chain Monte Carlo (MCMC) search methodology [2]–[5]. This technique applied to DT models has revealed promising results for real-world problems.

The Bayesian generalization of tree models that is required to evaluate the posterior distribution has been given in [2]. Recently, evaluating the posterior distribution of DTs has been suggested by using a reversible jump (RJ) MCMC technique [5]. The RJ MCMC technique itself was originally introduced by Green [8].

For interpretation of DT ensembles, two approaches have been suggested [9], [10]. The first is based on an idea of clustering DTs in a two-dimensional (2-D) space given by DT size and DT fitness. The second approach is based on using a DT of maximum *a Posteriori* (MAP) probability.

In this paper, we present a new approach to probabilistic interpretation of the Bayesian DT ensembles. This approach is based on the quantitative evaluation of uncertainty of the DTs, and allows experts to find a DT that provides a high-predictive accuracy and confident outcomes. To make the Bayesian averaging (BA) over DTs feasible in our experiments, we use the RJ MCMC technique described in [11]. The classification uncertainty is evaluated within an uncertainty envelope technique, dealing with the class posterior distribution and a given confidence probability, which we developed and described in [12]. Using this evaluation technique in our experiments, we find that in terms of the predictive accuracy, the proposed method outperforms the MAP method of interpreting DT ensembles. The comparisons are made on medical data sets, taken from the University of California at Irvine (UCI) Machine Learning Repository [13], as well as on the trauma data, which are mainly represented by the attributes of the trauma injury severity score (TRISS) model originally described in [14].

The above TRISS model is based on a multiple regression to estimate the probability of survival of a patient from the injury-severity score, revised trauma score, age, and type of injury (blunt and penetrating). The coefficients of this model have been estimated for the two types of injury and different ages.

Manuscript received May 17, 2005; revised November 1, 2005 and February 7, 2006. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under the Critical Systems Program, under Grant GR/R24357/01.

V. Schetinin is with the Computing and Information System Department, University of Bedfordshire, Luton LU1 3JU, U.K. (e-mail: vitaly.schetinin@beds.ac.uk).

J. E. Fieldsend, D. Partridge, W. J. Krzanowski, R. M. Everson, T. C. Bailey, and A. Hernandez are with the School of Engineering, Computer Science, and Mathematics, University of Exeter, Exeter EX4 4QF, U.K. (e-mail: j.e.fieldsend@exeter.ac.uk; d.partridge@exeter.ac.uk; w.j.krzanowski@exeter.ac.uk; r.m.everson@exeter.ac.uk; t.c.bailey@exeter.ac.uk; a.hernandez@exeter.ac.uk).

T. J. Coats is with the Accident and Emergency Department, Leicester Royal Infirmary, Leicester LE1 5WW, U.K. (e-mail: tc61@le.ac.uk).

Digital Object Identifier 10.1109/TITB.2006.885551

Obviously, such multiple regression models as the TRISS model are not convenient for interpretation purposes, failing to make the decision process understandable and transparent, which is desirable for experts.

Section II describes the basis of the Bayesian RJ MCMC technique of averaging over DTs, which is used in our experiments. Section III then describes our approach to the probabilistic interpretation of the Bayesian ensemble of DTs. Sections IV and V describe the experimental results obtained on real clinical data sets, and finally, Section VI concludes the paper.

II. Bayesian Averaging over Decision Trees

In a general classification problem, we wish to predict the class $(1,\ldots,C)$ of an m-dimensional data point $x = (x_1,\ldots,x_m)$ that is based on the values of the m predictors $x_1,\ldots,x_m,$, and we typically have available a set of training data D consisting of sets $(x_i, y_i), i = 1,\ldots,n$, where the categorical response $y_i \in \{1,\ldots,C\}$ gives the known class of each of n data points. The predictive distribution we are interested in can be written as an integral over parameters θ of the classification model as

$$p(y|x,D) = \int_\theta p(y|x,\theta)p(\theta|D)\,d\theta \tag{1}$$

where D denotes the given training data.

However, the integral (1) can be analytically calculated only in simple cases. In practice, part of the integrand in (1), which is the posterior density of θ conditioned on the data $D, p(\theta|D)$, cannot usually be evaluated exactly, but only to within a constant of proportionality. However, if values $\theta^1,\ldots,\theta^{(\mathrm{r})}$ are drawn from the posterior distribution $p(\theta|D)$, we can write

$$p(y|x,D) = \int p(y|x,\theta)p(\theta|D)d\theta \approx \frac{1}{R}\sum_{r=1}^{R} p(y|x,\theta^{(r)},D) \tag{2}$$

where R is the given number of samples.

This is the basis of the MCMC technique for approximating integrals [3]. To perform the approximation, we first need to draw samples of $\theta^{(r)}$ from $p(\theta|D)$. This is done by defining a Markov Chain on the parameter values $\theta^{(r)}$, with transition density from $\theta^{(r)}$ to $\theta^{(r+1)}$ given by $q(\theta^{(r+1)}|\theta^{(r)})$. Such a transition density can ofcourse be defined in many different ways, but if it is done according to the Metropolis–Hastings algorithm (see, e.g., [5, pp. 32–34]), then the stationary distribution of the Markov Chain is identical to the posterior distribution $p(\theta|D)$ that we wish to sample from. The definition of $q(\theta^{(r+1)}|\theta^{(r)})$ is specific to the classification problem, and we sketch its components for DTs below, but once it has been done, we need to run the Markov Chain from an arbitrary starting value $\theta^{(0)}$ until its output has converged to a stationary distribution—this phase is called the *burn-in*. In practice, the stationarity of distribution can be determined visually or quantitatively by using χ^2 tests. When a Markov Chain becomes stable, we can draw desired samples $\theta^{(r)}$ and calculate the predictive posterior density (2)—this phase is called the *post burn-in*.

Let us now turn to the specific case of DTs. The DT parameters are defined as $\theta = (s_l^{\mathrm{pos}}, s_l^{\mathrm{pred}}, s_l^{\mathrm{rule}})$, where l are the indices of spitting nodes in the tree, $s_l^{\mathrm{pos}}, s_l^{\mathrm{pred}}, \mathrm{and} s_l^{\mathrm{rule}}$ define the *position*, *predictor*, and *rule* for each splitting node, respectively. The priors for these parameters can be specified as follows. First, we can define a maximal number of splitting nodes, say, $s_{\max} = n - 1$, so $s_l^{\mathrm{pos}} \in \{1,\ldots,s_{\max}\}$. Second, we can draw any of the m predictors from a uniform discrete distribution $U(1,\ldots,m)$ and assign $s_l^{\mathrm{pred}} \in \{1,\ldots,m\}$. Finally, the candidate value for the predictor $s_l^{\mathrm{pred}} = p$ can be drawn from a uniform discrete distribution $U(q_p^{(1)},\ldots,q_p^{(M)})$, where $q_p^{(1)},\ldots,q_p^{(M)}$ is the given set of splitting rules for predictor s_l^{pred}, either categorical or continuous.

From graph theory (see, e.g., [15]), we know that there are $S_k = 1/k+1\begin{pmatrix}(2k\\k\end{pmatrix}$ possible ways of constructing binary DTs with k splitting nodes. This means that the number of the DT structures providing k splits can be very large, e.g., for $k = 20, S_k = 6.6 \times 10^9$. Only a few of these DT structures can provide the desired maximum of posterior density (2). Obviously, to find the desired DT structures within an acceptable computational time, we need to avoid the exhaustive search of all possible DT structures. Such avoidance is achieved when the search can be started with DTs containing one splitting node, and when new splitting nodes can be subsequently added to the DT while its posterior probability increases (see, e.g., [4], [5]). Clearly, for the reasonably large computational time, such a technique can find the suboptimal results.

Exploring the posterior probability of DTs, induced from real-world data has been suggested by using the following types of moves [4], [5].

- *Birth:* Randomly split the data points falling in one of the terminal nodes by a new splitting node with the predictor and rule drawn from the corresponding priors.
- *Death:* Randomly pick a splitting node with two terminal nodes and assign it to be a single terminal node with the united data points.
- *Change-split:* Randomly pick a splitting node and assign it a new predictor and rule drawn from the corresponding priors.
- *Change-rule:* Randomly pick a splitting node and assign it a new rule drawn from a given prior.

The first two moves, *birth* and *death*, are reversible and change the dimensionality of θ [8]. The remaining moves provide jumps within the current dimensionality of θ. Note that the *change-split* move is included to make "large" jumps, which potentially increase the chance of sampling from a maximal posterior, while the *change-rule* move does "local" jumps.

Because DTs are hierarchical structures, the changes at the nodes located at the upper levels can significantly change the location of data points at the lower levels. For this reason, there is a very small probability of changing and then accepting a DT located near a root node. Therefore, the MCMC algorithm tends to collect DTs in which the splitting nodes located far from a root node are changed. These nodes typically contain small numbers of data points. Subsequently, the value of log likelihood is not changed much, and such moves are usually accepted. As a result, the MCMC algorithm cannot explore a full class posterior distribution [4], [5]. However, the use of sweeping technique, making DTs shorter, allows us to obtain more accurate estimates of the posterior distribution [11].

III. Confident Interpretation of the Bayesian Decision Tree Ensembles

This section describes our method of interpreting Bayesian DT ensembles. First, we introduce the confidence in the classification outcomes of the DT ensemble, which can be quantitatively estimated on the training data. Then, we give an illustrative example of how a desired DT can be selected, and finally, we describe the main steps of our method.

TABLE I
Example of the Outcomes for the DT Ensemble

Data point	o_1	o_2	o_3	o_4	o_5	γ
$\mathbf{x}_1$	0	0	1	1	1	3/5
$\mathbf{x}_2$	0	1	1	1	1	4/5
$\mathbf{x}_3$	1	0	1	1	0	3/5
$\mathbf{x}_4$	1	1	1	0	1	4/5
$\mathbf{x}_5$	1	1	0	1	0	3/5
$\mathbf{x}_6$	1	0	0	0	1	2/5
$\mathbf{x}_7$	1	1	0	0	1	3/5
Sum	3	3	4	4	3	

A. Interpretation Strategy Using Classification Confidence

The idea behind our method of interpreting the Bayesian DT ensemble is to find within the ensemble a single DT that covers most of the training examples classified as confident and correct. For the DT ensemble, the confidence of classification outputs can be easily estimated by assessing the consistency of the classification outcomes [12]. Indeed, within a given classification scheme, the outputs of the ensemble depend on how well the classifiers were trained and how representative were the training data. For a given data point, the consistency of classification outcomes depends on how close this point is to the class boundaries. So for the c th class, the confidence in the ensemble can be estimated as a ratio γ_c between the number of classifier outcomes of the c th class, N_c, and the total number of classifiers $N: \gamma_c = N_c/N$, where $c = 1, \ldots, C$.

Clearly, the classification confidence is maximal, equal to 1.0, if all the classifiers assign a given data point to the same class, otherwise the confidence is less than 1.0.

The minimal value of confidence is equal to $1/C$ if the classifiers assign the given data point to the C classes in equal proportions. So for a given data point, the classification confidence in the ensemble can be properly estimated by the ratio γ.

Within the above framework in real-world applications, we can define a given level of the classification confidence $\gamma_0: 1/C \leq \gamma_0 \leq 1$, for which the cost of misclassification is small enough to be accepted. Then, for the given data point, the outcome of the ensemble is said to be *confident* if the ratio $\gamma \leq \gamma_0$. Clearly, on the labeled data, we can distinguish between *confident and correct* outcomes and *confident but incorrect* outcomes. The last outcome of the ensemble may appear due to noise or overlapping classes in the data. Otherwise, if the ratio $\gamma < \gamma_0$, then the outcome of the DT ensemble is declared to be *uncertain*.

Let us now consider how the above estimates of classification confidence can be used to interpret the Bayesian DT ensemble. Assume the following example with five classifiers and seven training examples $x_1, \ldots, x_7$, as presented in Table I. For the given data point x_i and the five classifiers, we can define five indicators $o_1, \cdots, o_5 : o_i = 1$ if $y_j = t_i$, otherwise $o_i = 0$, where y_j and t_i are the outcome of the j th classifier and the class label of data point x_i, respectively.

For each data point, the value of classifiers consistency γ was calculated; their values range between 4/5 and 2/5. Let a confidence level γ_0 be 3/5, above which the first five data points are classified as confident and correct, while the other two data points are classified as uncertain and confident but incorrect, respectively. Then, we can see that DT3 and DT4 cover a maximal number of data points, equal to four. Consequently, one of these DTs can be selected for interpreting the confident classifications. Such a selection can be reasonably done with a minimal DT-size criterion, because such DTs provide the best generalization ability.

In practice, the number of DTs in the ensemble as well as the number of the training examples can be large. Nevertheless, counting the number of confident and correct outcomes, as described in the above example, we can find a desired DT that can be used for interpreting the confident classification. The performance of such a DT can be slightly worse than that of the Bayesian DT ensemble, and Section V provides the experimental comparison of their performances. Next, the selection procedure is described.

B. Selection Procedure

Having an ensemble of DTs, we can find one or more DTs that cover a maximal number of the training examples classified as confident and correct, while the number of misclassifications on the remaining examples is kept minimal. To find such DTs, we can first select a set of DTs, S1, which cover a maximal number of the training examples classified by the DT ensemble as confident and correct under the given level γ_0. The part of the training data, which has been misclassified by the DT ensemble is then removed from the training data, and the remaining data, D1, are used to find among the set S1 a subset of DTs, S2, which provide a minimal error rate on the data D1. Finally, a DT of a minimal size is selected from set S2 to be assigned as the desired single DT. Thus, the main steps of the selection procedure are as follows.

Step 1) Among a given Bayesian DT ensemble, find a set of DTs, S1, which cover a maximal number of the training examples classified as confident and correct with a given confidence level γ_0.

Step 2) Find the training examples, which were misclassified by the Bayesian DT ensemble and then, remove them from the training data. Denote the remaining training examples as D1.

TABLE II
PREDICTORS DESCRIBING THE TRAUMA DATA

Predictor	Name	Type
x_1	Age	Continuous
x_2	Gender	$\{0, 1\}$
x_3	Injury (blunt or penetrating)	$\{0, 1\}$
x_4	Head injury	$\{0, 6\}$
x_5	Facial injury	$\{0, 4\}$
x_6	Chest injury	$\{0, 6\}$
x_7	Abdominal injury	$\{0, 5\}$
x_8	Limbs	$\{0, 5\}$
x_9	External injury	$\{0, 3\}$
x_{10}	Respiration rate	Continuous
x_{11}	Systolic Blood Pressure	Continuous
x_{12}	Glasgow Coma Score eye	$\{0, 4\}$
x_{13}	Glasgow Coma Score motor	$\{0, 6\}$
x_{14}	Glasgow Coma Score verbal	$\{0, 5\}$
x_{15}	Oximetry	Continuous
x_{16}	Heart rate	Continuous

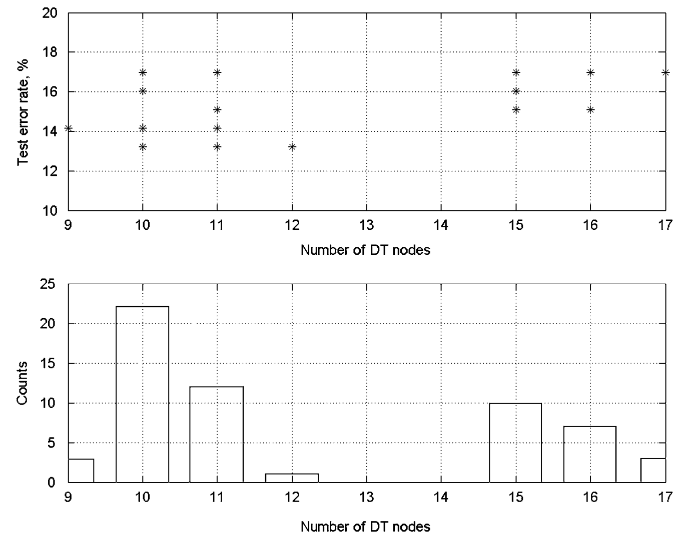

Fig. 1. (a) Testing error versus the number of DT nodes. (b) Distribution of 58 DTs over the numbers of their nodes.

Step 3) Among the set S1 of DTs, find those which provide a minimal misclassification rate on the data D1. Denote the found set of such DTs as S2.

Step 4) Among the set S2 of DTs, select those whose size is minimal. Denote a set of such DTs as S3. The set S3 contains the DTs any of them can be taken as the desired DT.

For the given example in Table I, the above procedure has selected DT3 and DT4, which cover a maximal number of the training examples, equal to four, classified as confident and correct with a given confident level $\gamma_0 = 3/5$. Let the DT3 and DT4 consist of 10 and 12 nodes, respectively, and put them in the set S1.

As the DT ensemble has misclassified the seventh data point, we remove this point and then put the remaining examples into the data set D1. Both DT3 and DT4 from the set S1 have misclassified two examples on the data set D1. Consequently, these DTs are allocated into the set S2. Finally, analyzing the sizes of the DTs, included in the set of S2, we select DT3 consisting of 10 nodes. Therefore, the set S3 contains only one DT3, which is assigned to be desired. Next, we discuss the use of the above selection procedure and compare the performance of the resultant DTs on some clinical problems.

IV. EXPERIMENTAL RESULTS

This section describes the experimental results obtained with the proposed technique on the trauma data collected in the Royal London Hospital.

A. Trauma Data

The trauma data used in our experiments consist of 316 labeled examples, which present the difficult cases for clinicians deciding on the survival of a patient [14]. These data have 16 predictors, as listed in Table II. This includes predictors such as *age, respiration rate, systolic blood pressure (BP)*, o*ximetry (%)*, and *heart rate*, which are continuous; the remaining predictors are nominal. Two hundred and ten data points randomly selected from the original data form a training data set, and the remaining 106 form a test data set. The proportions of surviving patients were 0.47 and 0.56 for the training and test data sets, respectively.

The Bayesian ensemble of DTs misclassified 11 training examples out of 210, and thus the training error was 5.24%. Among 5000 DTs collected during the post burn-in phase, we find 58 DTs, included into set S1, that cover all 123 confident and correct training examples. These DTs misclassified one example in the revised training data D1, containing $210 - 11 = 199$ examples.

For these 58 DTs, Fig. 1(a) and (b) show the test errors and distribution of DT nodes, respectively. We can see that the number of nodes in the DTs varies between 9 and 17, and that the DTs with 10 nodes were collected more frequently, i.e., 22 times.

From Fig. 1(b), we can see that there are three DTs consisting of a minimal number of nodes equal to nine, which misclassified 14.1% of the test data. All these DTs misclassified the same number of the test examples, and a DT randomly selected from this DT set provides the misclassification rate of 14.1% with a probability equal to $3/3 = 1.0$. The performance of the selected DT is worse than that obtained with the BA, only by 1.0% and so, in practice, this DT can be used for interpretation purposes.

In particular, observing Fig. 1(a), we can see that the misclassification rate of the DTs, on an average, slightly increases when the number of DT nodes increases. This happens because big DTs tend to overfit.

From Fig. 1(a), we can see that one or more DTs from the group of DTs, consisting of 10 nodes, provide a minimal test error rate of 13.2%. At the same time, a few DTs from this group can be reassigned to the set of DTs consisting of nine nodes, if we cut off a terminal node, which changes splitting the training data insignificantly. Note that such DT nodes appear because a prior information on the number of DT nodes has not been given in our experiments with the RJ MCMC technique (i.e., the

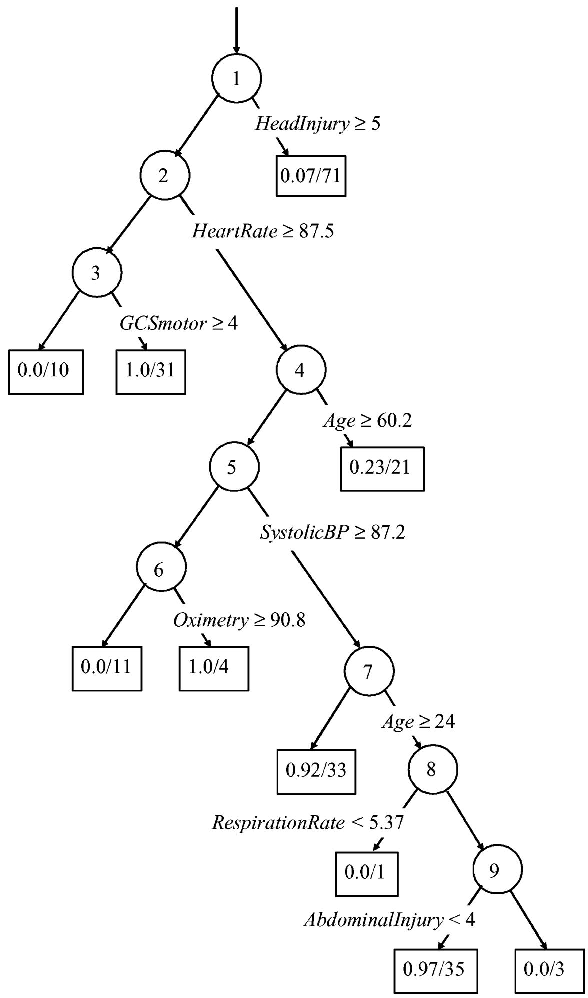

Fig. 2. Resultant DT consisting of nine nodes.

prior was "uninformative"). So the group of DTs, consisting of nine nodes and providing misclassification rate of 14.1%, can be enlarged by including the new reassigned DTs providing a better generalization ability. Obviously, this leads to increasing a chance of selecting a single DT that provides a minimal misclassification rate of 13.2%.

Alternatively to the above DT set, we can assume a DT, which is randomly selected from the largest group of DTs consisting of 10 nodes [see Fig. 1(b)]. In this case, the DT can be selected with the probability equal to 0.36, 0.05, 0.23, or 0.36 and the test error of 17%, 16%, 14.2%, or 13.2%, respectively. From this example, we can see that the alternative selection provides a DT of which the test error varies more widely than that obtained with the suggested selection procedure.

B. Interpretation of the DT

The resultant DT selected by the suggested procedure under $\gamma_0 = 0.99$ is presented in Fig. 2. This DT, originally consisting of 10 nodes, was obtained by cutting one splitting node. Each splitting node of the DT provides a specific question, which has a yes/no response, and two branches corresponding to these responses to the specific question; the positive response (yes) corresponds to the right-hand branch and the negative response (no) corresponds to the left-hand branch.

From Fig. 2, we can see that the first node asks question about *head injury*, the positive response which is associated with the terminal node splitting 71 training examples, with the probability of survival of a patient equal to 0.07. Note that node 8 asking *respiration rate* splits only one example of the training data, and under these circumstances this node can be removed.

As an explanation of the trauma decision process, this DT was judged to be physiologically plausible and a general fit with what would be expected from a clinical perspective. It seems to be picking out brain injury (*head score* and *Glasgow coma score* (GCS)), bleeding (*systolic BP* and *heart rate*), and preexisting physiological reserve (*age*) as important factors.

The main causes of death after injury are brain damage and bleeding. The early stage of the DT seems to be saying: if you have a severe head injury, it does not matter whether you are bleeding (reflected in physiological disturbance) or not, you are likely to die. If you do not have a severe head injury, the amount of physiological disturbance (bleeding or respiratory distress) and your capacity to respond to that disturbance becomes important. *Head injury*, the first splitting-node decision, fits with what we know about brain injury being a huge influence on the patient's outcome: even if you stop the bleeding, the patient will still die. It is interesting that there is a second group of patients that have a head injury score of less than five and a normal heart rate, where GCS motor response becomes important. This decision structure is suggestive of hypoxic brain injury; this hypothesis is the subject of further detailed medical examination.

With respect to Fig. 2, we note that the way the two age nodes are used is very interesting because current injury models [14] use only one with a cutoff at 55 years. We know that the extremes of age are very different in almost all areas of medicine, so the fact that there are two decision points, one for old and one for young, fits well with such preconceptions. The slight surprise is that there is not a younger cutoff, because preteen children have better outcomes, but there may be too few cases in this age range for the modeling process to identify this effect in the available data set.

Overall, it is interesting that little seems to be contributed by the "anatomical" attributes (i.e., the injury severity scores for each body area), apart from *head injury*. This suggests that in the U.K. there maybe a financial saving gained by eliminating the collection of less useful data, or that better results may be achievable by collecting more detailed data from the head region.

C. Cutoff Issue

Fig. 3 shows that 16 patients have *age* < 10 (i.e., preteens) in the training set, and seven in the test data. So, we might expect that with less than 10% of the training data in the preteen category, the surprisingly "old" cutoff of 24 (61 cases, which is more than 25%) may indeed reflect the paucity of preteen information in the training data.

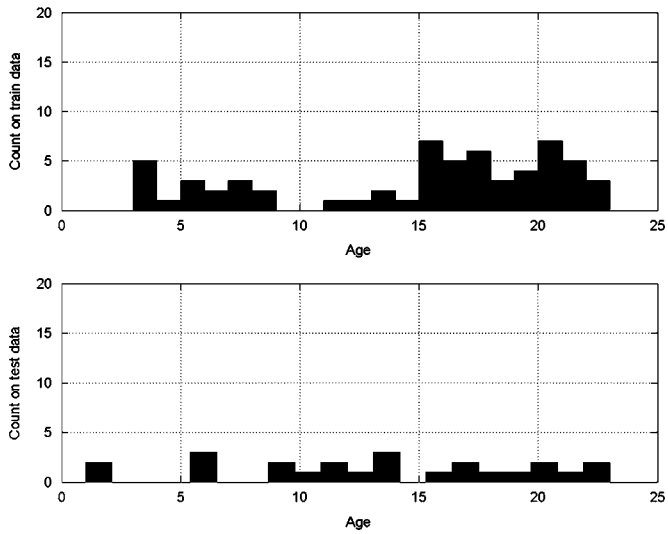

Fig. 3. Histograms of patient ages for the training and test data.

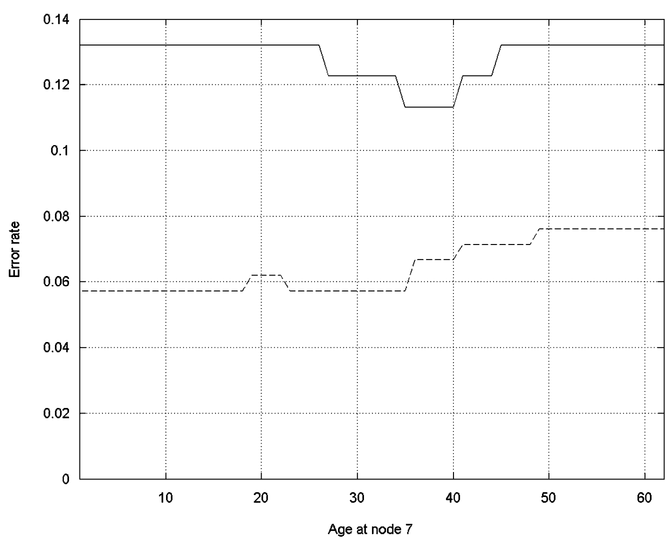

Fig. 4. Error rates of the DT versus *age* asked at node 7. The *solid line* is the test error and the *dashed line* is the training error.

We explored the significance of this age cutoff by retesting the DT, as presented in Fig. 2, with the latter age cutoff taking all integer values from 1 to 62 at node 7 asking *age*; Fig. 4 plots the training and test errors for this case. However, no clear performance gain was observed for a preteen cutoff value. In fact, classification accuracy was more or less constant from age 1 to 35, after which it decreased. In sum, we conclude that lack of sufficient patient records (see counts in Fig. 3) means that no confidence can be placed in the actual cutoff value selected.

Fivefold cross validation was used in the evaluation of predictor importance. The entire data set was split into five subsets of equal size, four of which were used for training and the remaining subset for testing. The results were averaged over five runs (using the different testing subsets).

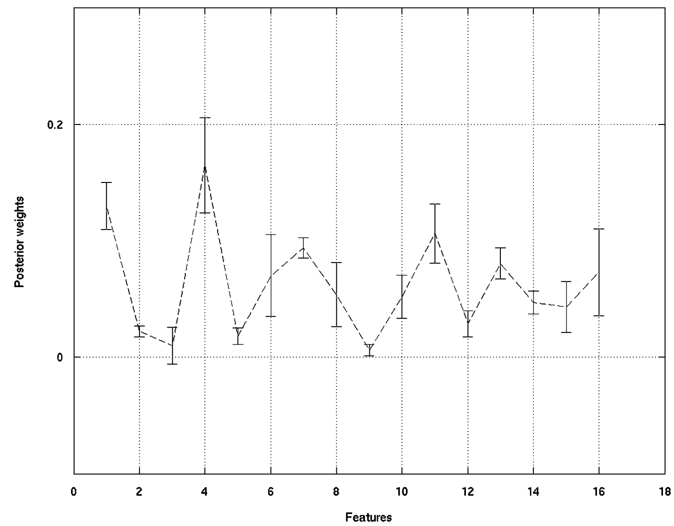

Fig. 5. Posterior weights of the predictors averaged within fivefold cross validation.

TABLE III
POSTERIOR WEIGHTS AND RANKS OF THE PREDICTORS

Predictor	Name	Posterior	Rank
x_4	Head Injury	0.165	1
x_1	Age	0.130	2
x_{11}	Systolic Blood Pressure	0.106	3
x_7	Abdominal Injury	0.094	4
x_{13}	Glasgow Coma Score Motor	0.081	5
x_{16}	Heart Rate	0.073	6
x_6	Chest Injury	0.070	7
x_8	Limbs	0.054	8
x_{10}	Respiration Rate	0.052	9
x_{14}	Glasgow Coma Score Verbal	0.047	10
x_{15}	Oximetry	0.043	11
x_{12}	Glasgow Coma Score Eye	0.029	12
x_2	Gender	0.022	13
x_5	Facial Injury	0.018	14
x_3	Injury (blunt or penetrating)	0.009	15
x_9	External Injury	0.006	16

In Fig. 5, we can see that predictors such as *age* (x_1), *head injury* (x_4), *abdominal Injury* (x_7), *systolic BP* (x_{11}), and *GCS motor* (x_{13}) are used in the Bayesian DTs, on average more frequently than the others. In contrast, predictor *external injury* (x_9) is used with a less frequency. Additionally, we can see from the error bars that the weighted posterior values (or posterior weights) of some predictors (e.g., for *head injury* (x_4), *chest injury* (x_6), and *heart rate* (x_{16})), have a high variance. Such wide deviations may be caused by variations in the training data within the five fold cross validation.

Table III lists the average posterior weights of the predictors and their ranks. Note that the rank of a predictor corresponds to the predictor index in a series sorted on the values of posterior weights, so that the ranks for the predictors with the maximal and minimal values of posterior weights are equal to 1 and 16, respectively. The smaller the rank of predictor, the bigger is its contribution to the classification.

TABLE IV
DATA SETS CHARACTERISTICS

	Pima	Wisconsin	Trauma1	Trauma2
n	768	683	316	1468
m	8	9	16	18

Trauma care is an area of medicine where there is an existing predictive model, and the factors influencing the probability of survival are relatively well understood [14]. There is a good correlation between the highly ranked factors in Table III and the factors that clinicians regard as important for their patients.

Brain injury tends to have more complications, so outcome is directly and strongly related to the extent of brain injury, as seen in the high rankings for head injury and GCS motor (which is known to be the most reliable of the three components of the GCS). The ability of the body to cope with injury is directly related to age (for example, there is a rough rule of thumb in burns patients that if the percentage body area burnt plus age exceeds 100, the patient will die). The importance of age is seen as its high rank.

It is interesting that the blunt/penetrating distinction does not have much influence as this is an important factor in the conventional predictive model, which originates from U.S. databases. However, in this U.K. data set there will be many fewer patients with penetrating injury (usually shooting or stabbing), so it may not actually be an important predictor in the U.K. population.

Respiratory rate and *oximetry* are ranked rather lower than might be expected; this is an interesting area for further exploration. Are our clinical perceptions wrong, or are patients with respiratory distress poorly represented in this data set, or are the respiratory measures so closely correlated with some other factor(s) that they add little predictive weight?

V. COMPARISON OF TECHNIQUES

In this section, we compare our technique of extracting a confident decision tree (CDT) with the BA and the MAP techniques described in [10]. The comparison is made in terms of misclassification rate within fivefold cross validation on two medical data sets known as Pima and Wisconsin [13]. Trauma 1 are the data described in Section IV-A. Trauma 2 is another set of the trauma data collected at the Royal London Hospital; 23% of them represent patients who died. For all these experiments, the confidence level γ_0 was given equal to 0.99. All the data sets are two-class problems. The number of the labeled examples n and the number of predictors m are listed in Table IV.

The Bayesian DT ensemble technique ran with the pruning factor $p_{\min}$ (a minimal number of data points allows to be in splits) ranging between 5 and 50 for Pima, Wisconsin, and Trauma 2, while for the Trauma 1, $p_{\min} = 1$. The number of burn-in and post burn-in samples, and sampling rates were set to 10 000, 5000, and 7, respectively. Note that under the given sampling rate, during post burn-in every seventh sample is collected. This allows the independence of samples to increase, and subsequently to improve the result of model averaging [5].

The proposal probabilities with which the birth, death, change-split, and change-rule are made during the MCMC search were set to 0.1, 0.1, 0.2, and 0.6, respectively. The proposal distribution for the change moves was set a Gaussian $N(0, \sigma)$ with variance $\sigma = 1.0$. The misclassification rates of the above three techniques, BA, CDT, and MAP, are shown in Table V. Clearly, in the theory, the BA technique should provide fewer misclassifications than the CDT and MAP techniques. In our experiments, however, we can observe that all these techniques have nearly the same misclassification rates within fivefold cross validation. Nevertheless, comparing the average rates of misclassification, we can see that the BA and CDT techniques slightly outperform the MAP technique. At the same time, the average misclassification rates of the BA and CDT techniques are almost the same.

TABLE V
MISCLASSIFICATION RATES OF DTS ON THE TEST DATA SETS WITHIN FIVEFOLD CROSS VALIDATION

	Pima	Wisconsin	Trauma1	Trauma2
BA	27.7±6.5	3.8±4.5	**13.1**±7.5	**15.5**±3.7
CDT	**26.7**±7.5	**3.2**±3.9	14.1±7.7	16.1±4.4
MAP	27.8±8.4	3.9±2.8	18.5±2.6	17.1±3.4

TABLE VI
SIZES OF DTS WITHIN FIVEFOLD CROSS VALIDATION

	Pima	Wisconsin	Trauma1	Trauma2
BA	17.8±1.1	11.3±1.8	12.5±1.0	**41.2**±3.7
CDT	**17.6**±3.9	**9.6**±6.6	**10.6**±3.6	42.8±11.9
MAP	21.7±2.4	14.2±6.5	16.8±1.7	53.2±5.5

TABLE VII
DECREASES IN TEST ERROR AND DT SIZE FOR THE CDT TECHNIQUE WITH RESPECT TO THE MAP TECHNIQUE

	Pima	Wisconsin	Trauma1	Trauma2
Test error, %	1.18	0.74	4.22	0.88
Size	4.10	4.60	2.60	10.40

Table VI shows the sizes of DTs induced in our experiments within fivefold cross validation. From this table, we can certainly conclude that the CDT technique provides shorter DTs than the MAP technique.

Table VII provides the estimates of decreases in the test error as well as in the DT size for the CDT technique with respect to the MAP technique. The values of these estimates were averaged within the fivefold cross validation.

Figs. 6 and 7 provide the comparisons of the CDT and MAP techniques on the four data sets over fivefold cross validation. The comparisons are made in terms of the test error and the DT sizes, respectively.

The above comparison allows us to conclude that the CDT technique outperforms the MAP technique in terms of misclassification error on the test data and, especially, in the size of DTs.

VI. CONCLUSION

DTs, particularly when set within a framework of BA, prove to be powerful automatic classification systems, which in the trauma domain, at least, outperform the traditional selection of decision structures in terms of classification uncertainty [10]. In

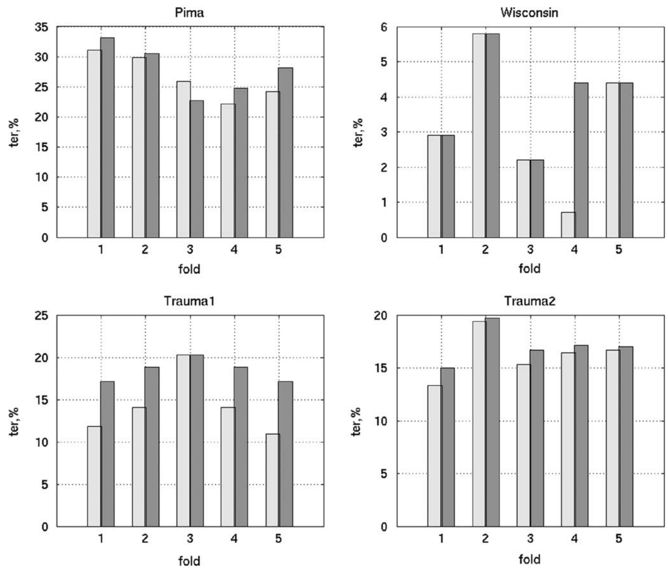

Fig. 6. Misclassification rates of the DTs on the test data sets obtained by the CDT (*gray bars*) and by the MAP (*dark bars*).

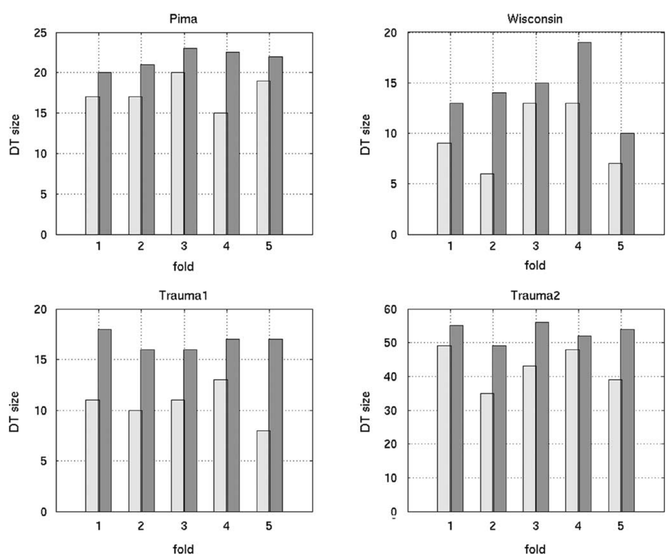

Fig. 7. Sizes of the DTs obtained by the CDT (*gray bars*) and by the MAP (*dark bars*).

addition, a BA approach offers the possibility of an estimate of the confidence to be attached to every prediction.

However, perhaps more importantly DT classifiers are said to be preferred (in contrast to, say, neural net classifiers) because they are interpretable, and this property will facilitate the use of DT models to extract useful knowledge about the optimal decision processes within the application domain. The biological plausibility of DTs may well be more acceptable to clinicians than a "black box."

Because the Bayesian ensemble of DTs does not appear to be a sensible concept with respect to interpretability of the optimal decision processes, a selection procedure for extracting confident DTs from the Bayesian DT ensemble was proposed and demonstrated. A selected tree was judged to be usefully explanatory of the trauma decision process. Objective evidence for useful explanatory power was provided in terms of both a subsequent focus of attention on specific features (e.g., the age cutoff) that resulted in the extraction of new knowledge about the role of this predictor, and objective confirmations of the roles of certain other predictors.

REFERENCES

[1] R. O. Duda and P. E. Hart, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2001.
[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
[3] W. Buntine, "Learning classification trees," *Statist. Comput.*, vol. 2, pp. 63–73, 1992.
[4] H. Chipman, E. George, and R. McCullock, "Bayesian CART model search," *J. Amer. Statist. Soc.*, vol. 93, pp. 935–960, 1998.
[5] D. Denison, C. Holmes, B. Malick, and A. Smith, *Bayesian Methods for Nonlinear Classification and Regression*. New York: Wiley, 2002.
[6] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. New York: Wiley, 2004.
[7] J. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[8] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.
[9] H. Chipman, E. George, and R. McCulloch, "Making sense of a forest of trees," in *Proc. 30th Symp. Interface*, 1998, vol. 29, pp. 84–92.
[10] P. Domingos, "Knowledge discovery via multiple models," *Intell. Data Anal.*, vol. 2, pp. 187–202, 1998.
[11] V. Schetinin, J. E. Fieldsend, D. Partridge, W. J. Krzanowski, R. M. Everson, T. C. Bailey, and A. Hernandez, "The Bayesian decision tree technique with a sweeping strategy," presented at the 2004 Int. Conf. Advances Intell. Syst.—Theory Appl. Cooperation IEEE Comput. Soc., Luxembourg.
[12] J. E. Fieldsend, T. C. Bailey, R. M. Everson, W. J. Krzanowski, D. Partridge, and V. Schetinin, "Bayesian inductively learned modules for safety critical systems," in *Proc. 35th Symp. Interface Comput. Sci. Statist.*, vol. 35, Salt Lake City, UT, 2003, pp. 110–125.
[13] C. L. Blake and C. J. Merz. (1998). UCI Repository of machine learning data set. Univ. California Irvine, [Online]. Available: www.ics.uci.edu/~mlearn/MLRepository
[14] C. R. Boyd, M. A. Tolson, and W. S. Copes, "Evaluating trauma care: The TRISS method," *J. Trauma*, vol. 27, pp. 370–378, 1988.
[15] D. B. West, *Introduction to Graph Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2000.

Vitaly Schetinin (M'05), photograph and biography not available at the time of publication.

Jonathan E. Fieldsend (S'00–M'02), photograph and biography not available at the time of publication.

Derek Partridge, photograph and biography not available at the time of publication.

Timothy J. Coats, photograph and biography not available at the time of publication.

Wojtek J. Krzanowski, photograph and biography not available at the time of publication.

Richard M. Everson, photograph and biography not available at the time of publication.

Trevor C. Bailey, photograph and biography not available at the time of publication.

Adolfo Hernandez, photograph and biography not available at the time of publication.