Journal of Data Science **10**(2012), 61-73

Stability and Structure of CART and SPAN Search Generated Data Partitions for the Analysis of Low Birth Weight

Roger J. Marshall[1]* and Panagiota Kitsantas[2]
[1] *University of Auckland and* [2] *George Mason University*

Abstract: Searching for data structure and decision rules using classification and regression tree (CART) methodology is now well established. An alternative procedure, search partition analysis (SPAN), is less well known. Both provide classifiers based on Boolean structures; in CART these are generated by a hierarchical series of local sub-searches and in SPAN by a global search. One issue with CART is its perceived instability, another the awkward nature of the Boolean structures generated by a hierarchical tree. Instability arises because the final tree structure is sensitive to early splits. SPAN, as a global search, seems more likely to render stable partitions. To examine these issues in the context of identifying mothers at risk of giving birth to low birth weight babies, we have taken a very large sample, divided it at random into ten non-overlapping sub-samples and performed SPAN and CART analyses on each sub-sample. The stability of the SPAN and CART models is described and, in addition, the structure of the Boolean representation of classifiers is examined. It is found that SPAN partitions have more intrinsic stability and less prone to Boolean structural irregularities.

Key words: Boolean expression, CART, classification, stability, SPAN.

1. Introduction

Classification and regression tree (CART) analysis is increasingly popular, especially in public health and clinical medicine, to determine levels of risk among population sub-groups (Zhang *et al.*, 1996; Zhang and Singer, 1999). Specific examples include studies conducted to predict adherence to treatment for drug use (Dean *et al.*, 2009), growth in adolescents (Pawloski and Kitsantas, 2008), smoking among youth (Kitsantas *et al.*, 2007) and pre-term delivery (Zhang and Bracken, 1995). Also for identifying risk factors for low birth weight babies Kitsantas *et al.*, 2006), which is the impetus for the present study.

*Corresponding author.

Although CART is useful to uncover complex relationships among a large set of variables analysis, we consider two criticisms of trees. First, that tree classifiers can be unstable because of their hierarchical nature; all splits depend on previous splits, so that the tree obtained from one realisation a data set may differ markedly from that of another realisation. Tree instability here is thought of as a measure of sampling variability of tree structures. A second criticism is that Boolean structures generated by CART can also be hard to interpret because successively splitting on variables necessarily induces mutually exclusive sub-groups that are defined by awkward Boolean conjunctions of positive and negative attributes, that is, when a feature that is considered positively associated is combined with one that is negatively associated.

Several studies have proposed methods to alleviate the instability issue in classification trees (Breiman, 1996; Breiman, 1998; Dannegger, 2000). To our knowledge, however, there is little about how the stability of CART compares to other, non-hierarchical, classification techniques. One such method is search partition analysis (SPAN) (Marshall, 2006). In SPAN, a search is made for a partition that is in some sense optimal from class of Boolean partitions. The search is not done as in CART, by successive local searches at each level of hierarchical subdivision, but by applying partitions to the whole data set, and implementing a global search. Therefore SPAN seems less likely to be affected by the instability problem of tree growing. However, this remains speculative and in this study we offer an empirical investigation to examine it. A further feature of SPAN is that the class of Boolean combinations that is searched precludes the possibility of conjunctions of positive and negative attributes, so avoiding the second criticism of trees.

Although the instability problem of trees is well known (Dannegger, 2000), the degree of instability, in terms of the extent to which Boolean structures differ, seems not to have been explicitly addressed. Further, when nodes of trees are joined to form a classifier the Boolean structure often simplifies, possibly with some redundant terms (Marshall, 2001). This aspect of tree based partitions is also not often considered; trees are usually reported as trees rather than by simplified Boolean expressions derived from them. The purpose of this paper is to examine how Boolean structures of trees grown on data subsets differ, after simplification, and to compare them with corresponding partitions from a SPAN analysis.

To achieve this we focus on comparison of SPAN and CART analyses to identify mothers at high risk for having a low birth weight infant, using data from Florida (Kitsantas *et al.*, 2006). The data set is large, about 180,000 mothers. We randomly divide it into ten sub-samples, each of about 18,000 mothers, and perform ten separate SPAN and CART analyses. Since the sample and sub-

sample sizes are large, some degree of stability might be expected, in the sense that identical or somewhat similar partitions would be generated on each of the ten sub-samples. The results of the partitions generated by SPAN and by CART are presented.

2. Methods

2.1 Florida Birth Weight Data

Data comprised about 180,000 routine births in 1998 from the State of Florida with a binary outcome variable indicating low birth weight ($\leq$ 2500 grammes). The following variables were considered as predictors in these analyses: mother's age, ethnicity, smoking, marital and educational status, parity, weight gain during pregnancy, and adequacy of prenatal care. Table 1 gives the codings and categorisations of these variables. More details of the data can be found elsewhere (Kitsantas *et al.*, 2006). We randomly divided the data into ten non-overlapping samples of equal sizes, about 18,000 complete observations in each, and separately analysed each sub-sample by both CART and SPAN, applying precisely the same criteria (as described below) to each sub-sample.

Table 1: Coding in the data set and the codes and combinations used to define positive attributes, that is, attributes associated with low birth weight, with symbols used for their representation (in Tables 2 and 3)

Variable	Coding	Positive attributes code combinations
race	0 = black, 1 = Hispanic, 2 = Other	$B = 0$, $H = 1$, $O = 2$,
	3 = White	$B_h = 01^*$, $B_o = 02$, $B_{ho} = 012$
married	0 = no 1 = yes	$U = 0$
level of education	0 = elementary, 1 = some high school,	$E = 0$, $E_2 = 01$,
	2 = high school, 4=college, 3=higher	$E_3 = 012$, $E_4 = 0124$
age	0 < 18, 2 = 18-34, 1 > 34	$A = 0$, $A_2 = 02$
smoking tobacco	0 = yes, 1 = no	S=0
weight gain	0 < 20 lbs, 2 = 20-40lbs, 1 > 40lbs	$W = 0$, $W_2 = 02$
medical problem	0 = yes, 1 = no	$M = 0$
parity	0 = no previous babies, 2 = one previous, 1 => 1	$P = 0$, $H = 1$ $P_H = 01$
prenatal care	0 = inadequate, 1 = adequate	I=0

* e.g. $B_h = 01$ means either black or Hispanic

2.2 CART

Classification tree construction using the procedures proposed by Breiman has been summarily described in many sources (Carmelli *et al.*, 1997; Zhang *et al.*, 1996; Zhang and Bracken, 1995; Nelson *et al.*, 1998; Kitsantas *et al.*, 2006;

Zhang and Singer, 1999). The analyses reported here were done with the commercial CART software package (CART, 2000) that is based on the Breiman *et al.* (Breiman *et al.*, 1984) methodology. Here we used the Gini diversity splitting criterion, with assumed equal priors and equal misclassification costs. Typically the optimal trees, as determined by cross-validated minimum misclassification, had between 4 and 15 nodes. Our analyses were based on the marginally suboptimal trees with at most 6 terminal nodes. Each split was based on a binary attribute, say X_i, "going to the right" and its complement $\overline{X}_i$ "going to the left".

In CART, each terminal node is assigned to a prediction class and for a binary outcome combining the "class 1" nodes defines a binary partition of the sample space $\{\overline{A}, A\}$. The terminal node j represents a Boolean conjunction of attributes, C_j. For example, a terminal node after three splits of the hierarchy might be represented by the conjunction of attributes $C_j = X_2 \cap X_4 \cap \overline{X}_3$. This is the pathway tracing branches from an initial split on X_2 going right, next split on X_4 also going right, and finally split on X_3 going left. The A-side of the binary partition is the union of all the "class 1" terminal nodes. Formally, if T_1 is the set of class 1 and T_0 the set of class 0 terminal nodes

$$A = \bigcup_{j \in T_1} C_j \quad \text{and} \quad \overline{A} = \bigcup_{j \in T_0} C_j,$$

each of A and $\overline{A}$ defining a "disjunctive normal form (dnf)". The elements, or clauses, Cj induce mutually exclusive subgroups of the data, since each is a terminal node of the tree.

The Boolean expressions for A and $\overline{A}$ are typically not in their simplest form. For the purpose of comparing the trees expressions with the SPAN partitions, which are generated as dnfs in their simplest form, the CART expressions were also reduced to their simplest dnf by the rules of Boolean algebra.

2.3 SPAN

SPAN (Marshall, 2006) is also a search methodology for an optimal binary partition of the sample space into $\{\overline{A}, A\}$. As for CART, the two "sides" of a partition, are represented by Boolean expressions. An essential feature of a SPAN partition, however, is that $X_1, X_2, \cdots$ are defined a priori to be "positive" with respect to the outcome variable and mixtures of positive and negative attributes on the A or $\overline{A}$ sides of the partition are precluded, enabling a search strategy to be feasible. Partitions of this nature have been called "regular" (Marshall, 1986). Usually, there is a known direction of an association, so that labelling attributes positive is not restrictive and imposes a natural structure. Formally a SPAN

partition is of the form

$$A = \bigcup_{j=1}^{q} S_j, \tag{2.1}$$

where the S_j are q conjunctive clauses of positive attributes. The complement, expressed in dnf, is of the form

$$\overline{A} = \bigcup_{j=1}^{r} S'_j, \tag{2.2}$$

where the S'_j are r conjunctions of negative attributes $\overline{X}_1, \overline{X}_2, \cdots$. A set of m positive attributes, $T_m = \{X_1, X_2, \cdots, X_m\}$, is used as a pool of attributes for the search, which is done by initially generating all possible partitions of the form of (2.1) and (2.2) with restrictions on q and r and maximum length of S_j and S'_j (default ≤ 2) to make a search feasible. The most effective partition is then added into T_m, which becomes size $m + 1$, and the search begins anew in an iterative procedure.

Formally the algorithm is:

Step 1. Specify m (usually ≤ 12) and establish a set T_m, usually by ranking a larger set of attributes according to some measure of discrimination, say G (here Gini index of diversity).

Step 2. Exhaustively generate all possible partitions of the form (2.1) and (2.2) subject to $q \leq Q$ and $\text{card}(S_j) \leq Q$ (default $Q = 2$) from T_m. This exhaustive search is done using a "lock and key" algorithm (Marshall, 2000). Select the partition with largest G from this search, say $\{\overline{a}_1, a_1\}$.

Step 3. Augment T_m with a_1 to become $T_{m+1} = \{T_m, a_1\}$.

Step 4. Exhaustively generate all possible partitions of the form (2.1) and (2.2) subject to $q \leq Q$ and $\text{card}(S_j) <= Q$ from T_{m+1}. This exhaustive search is done adapting the "lock and key" algorithm to exclude possibilities already done in Step 2. Select the partition with largest G from this search, say $\{\overline{a}, a\}$.

Step 5. If $a = a_1$ terminate the search with $\{\overline{a}, a\}$ the optimal partition. Otherwise set $a_1 = a$ and return to Step 3.

Note that Step 1 requires specifying the direction of the relationship and labelling an attribute as positive. This may be a user-specified judgement or

determined from the data. Also Step 1 may include, where a predictor is non-binary, determining optimal combinations to define X_i. For example, the best cut-off for continuous predictors, or combinations of categorical states.

Boolean algebraic simplifications of generated partitions are continually done so that at each step a is expressed in terms of the primary elements of T_m. For example, suppose, at Step 4 of the cycle $a_1 = (X_1X_3)(X_1X_4)$ and $(a_1X_5)(X_3)$ is a generated partition. Substituting for a_1 gives $(X_1X_3X_5)(X_1X_4X_5)(X_3) = (X_1X_4X_5)(X_3)$.

Also the criterion to assess a partition's effectiveness, G, is complexity penalised (Marshall, 1995) to ensure partitions remain parsimonious. Iterations usually converge in two to four cycles. Here complexity is defined as $c = q + r - 1$ when A and $\overline{A}$ are expressed in their simplest disjunctive normal forms as in equations (2.1) and (2.2).

We used, as for the CART splitting criterion, the Gini diversity index, with equal priors, and equal misclassification costs to assess partition effectiveness. Initially the "cutoff" or combination of positive classes for each variable was established (Step 1 of the algorithm) and an attribute set of $m = 9$ was established, one for each of the 9 variables. These are shown in Table 1. For underlying interval measures (age, weight gain, education years) "positive" was categorised as below a given cut off since low values are known to be associated with low birth weight. For the race variable, we combined categories of the ethnic minorities. For parity, mothers having a first child are at a higher risk for LBW compared to those who have had more than one.

2.4 Instability Measures

Two partitions A_1 and A_2 are identical if $A_1 \cap A_2 = A_1 \cup A_2$ so that a measure of the difference between $A_1 \cup A_2$ and $A_1 \cap A_2$, provides a measure of dissimilarity. Extending this idea to the ten partitions A_i, $i = 1, \cdots, 10$ generated for each subsample, we consider the extent to which $U = \cup_i A_i$ and $I = \cap_i A_i$ differ. They may differ in two respects: in terms of how they partition the data and in terms of the structure of the Boolean expressions they represent. In terms of partitions of the data space, a stability index, Q can be constructed:

$$Q = \frac{\#(I)}{\#(U)},$$

where here # denotes the number of objects (mothers) in the set. Q, expressed as a percentage, is 100% when all data partitions precisely coincide, and there is full stability, and 0% when there is no mother who appears in all ten A_i, that is, I is empty.

Assessing whether the Boolean expressions for the partitions are similar in structure, in terms of their Boolean arrangement, is done by simply noting the commonalities of the conjunctive clauses of the dnf representation of the partitions.

3. Results

Table 2 shows the representation of the CART trees generated on each for the ten sub-samples. We give the representation of the "A-side", that is, low birth weight class. The partition of each tree in its full form is given and its representation after Boolean simplification. From the full form, it is clear that in most cases the initial split is on either M or W, in both cases resulting in an immediate terminal node. However, of the ten generated trees, none are the same. Seven are, after simplification, regular, but three are not (samples 6, 7 and 9) and possess combinations of positive and negative attributes. For example, the combination $\overline{M}B_o$ for sub-sample 6, not having a medical condition in conjunction with being "Black or other" constitutes one of them.

Table 2: CART representation of partitions for high of low birth weight built on each sub-sample. Showing full form by tracing tree branches and simplified forms. For brevity the $\cup$ and $\cap$ operators are omitted, that is, $(X \cap Y) \cup (Z) = (XY)(Z)$. $|T|$ is number of terminal nodes. * indicates that a partition is irregular

Sub-Sample	Full form: terminal nodes	$\lvert T\rvert$	Simplified form
1	$(W)(\ \overline{W}M)(\ \overline{W}\ \overline{M}B_oE_4)$	5	$(W)(M)(B_oE_4)$
2	$(M)(\ \overline{M}B)(\ \overline{M}\ \overline{B}S)(\ \overline{M}\ \overline{B}\ \overline{S}WP)$	6	$(M)(B)(S)(WP)$
3	$(M)(\ \overline{M}W)(\ \overline{M}\ \overline{W}P_H)$	5	$(M)(W)(P_H)$
4	$(W)(\ \overline{W}M)(\ \overline{W}\ \overline{M}B)$	4	$(W)(M)(B)$
5	$(M)(\ \overline{M}B_o)(\ \overline{M}\ \overline{B}_oS)(\ \overline{M}\ \overline{B}_o\ \overline{S}WE_2)$	6	$(M)(B_o)(S)(WE_2)$
6	$(W)(\ \overline{W}MB_h)(\ \overline{W}\ \overline{M}B_o)(\ \overline{W}M\ \overline{B}_hP_H)$	6	$(W)(MB_h)(\ \overline{M}B_o)(MP_H)^*$
7	$(M)(\ \overline{M}W)(\ \overline{M}\ \overline{W}B_oE_3)$	5	$(M)(W)(B_oE_3)$
8	$(M)(\ \overline{M}WU)(\ \overline{M}W\ \overline{U}P)(\ \overline{M}\ \overline{W}B)$	6	$(M)(WU)(WP)(\ \overline{W}B)^*$
9	$(M)(\ \overline{M}W)(\ \overline{M}\ \overline{W}UB)(\ \overline{M}\ \overline{W}U\ \overline{B}\ \overline{A}_2)$	6	$(M)(W)(UB)(U\ \overline{A}_2)^*$
10	$(W)(\ \overline{W}MU)(\ \overline{W}M\ \overline{U}W_2)$	5	$(W)(MU)(MW_2)$
Whole sample	$(M)(\ \overline{M}W)(\ \overline{M}\ \overline{W}B_oP_H)$	5	$(M)(W)(B_oP_H)$

Table 3 shows the SPAN partitions for each sub-sample and puts them against the simplified tree CART generated partitions. For the CART partitions, as already noted, no two trees generated the same partition; all the partitions are unique. However, for SPAN, there are two partitions that appear twice (for sub-samples 8 and 10, and samples 2 and 6).

Table 3: SPAN and CART representation of partitions for high risk of low birth weight built on each sub-sample and the whole sample

Sub-Sample	SPAN	Simplified CART form (from Table 2)
1	$(W)(M)(B_o)$	$(W)(M)(B_oE_4)$
2	$(M)(B)(W)(E_3S)$	$(M)(B)(S)(WP)$
3	$(M)(W)(B)(S)$	$(M)(W)(P_H)$
4	$(W)(M)(B)(E_3A)(E_3S)$	$(W)(M)(B)$
5	$(M)(B)(S)(WP)$	$(M)(B_o)(B_oS)(WE_2)$
6	$(M)(B)(W)(E_3S)$	$(W)(MB_h)(\overline{M}B_o)(MP_H)$
7	$(W)(B)(S)(MP)$	$(M)(W)(B_oE_3)$
8	$(M)(W)(B)$	$(M)(WU)(WP)(\overline{W}B)$
9	$(M)(W)(S)(BP)(BU)$	$(M)(W)(UB)(U\ \overline{A}_2)$
10	$(M)(W)(B)$	$(W)(MU)(MW_2)$
Whole sample	$(M)(W)(B)(S)$	$(M)(W)(B_oP_H)$

Table 4 shows the occurrences of common clauses of the partitions. For SPAN there were 11 unique conjunctive clauses in the dnf representations; for CART there were 20. The data stability index of the SPAN partitions is $Q = 57.6\%$ and for those by CART it is $Q = 38.8\%$, indicating greater stability for SPAN in terms of data partitions.

Table 4: SPAN and CART occurrences of the conjunctive clauses in dnf representation of the partitions

Common clauses	# partitions with clause	
	SPAN	CART
(M)	9/10	8/10
(W)	9/10	6/10
(B)	8/10	2/10
(S)	4/10	1/10
(E_3S)	3/10	-
(WP)	1/10	1/10
(B_o)	1/10	1/10
Other unique clauses	4	14
Total unique clauses	11	20

Table 5 gives the estimates of misclassification of the partitions and reduction in diversity when the partitions are applied to the entire data sample of 180,000 mothers. Clearly, in terms of these indices, there is little to choose between any of the SPAN and CART partitions. Finally, we note that there are no cases where the SPAN and CART partitions are the same.

Table 5: Complexity of SPAN and CART partitions. Also misclassification and reduction in Gini diversity, when applied to entire sample

Sample	Complexity		Misclassified %		Gini diversity reduction	
	SPAN	CART	SPAN	CART	SPAN	CART
1	3	4	36.62	36.52	.0370	.0374
2	5	5	36.27	36.15	.0401	.0398
3	4	4	36.54	36.89	.0398	.0344
4	6	3	36.27	36.41	.0401	.0378
5	5	5	36.15	36.92	.0398	.0357
6	5	5	35.96	36.49	.0404	.0368
7	5	4	35.80	36.40	.0411	.0373
8	3	5	36.41	36.47	.0377	.0368
9	6	5	36.16	36.30	.0404	.0379
10	3	4	36.25	37.31	.0381	.0331

4. Conclusions

Considering the large size of the sub-samples, about $n = 18,000$ in each and generated by random division of the pool of 180,000, it is perhaps surprising the extent to which partitions generated by both SPAN and CART on each sub-sample differ. None of the ten CART partitions are the same, and for SPAN there are eight different partitions. Despite the apparent instability, the prediction error and partition diversity is more or less the same for all partitions, and for both SPAN and CART, which probably indicates that the precise arrangement of key predictors does not, anyway, matter too much; there is no optimal partition, but a relatively flat response surface in the predictor space for this example. This seems likely to be the case in most real situations characterised by substantial diversity and by predictors that are generally weakly associated with outcomes, as is often the case in medicine, and low birth weight in particular.

As anticipated (see Introduction), SPAN partitions are rather more stable than CART. To explain this, in SPAN a "model" is fitted by applying it to the data as a whole, and although it is an iterative process in which the partition generated on the first cycle may alter the course of those subsequent, it is not applied piecemeal to the data. A tree, however, is fitted by a sequence of searches among ever smaller data subgroups, each is fixed by the previous split. The structure of a tree is fixed to a great extent by the initial split and although trees based on our data subsets often started out in the same way, with a split on medical problems, they quickly tended to diversify as the tree grows.

There are some limitations of the study. First, that it is a limited comparison, between just two methods. One is a method (SPAN) developed by one of us (RJM, e.g. Marshall, 2006), the other a commonly used tree growing algorithm.

Future work might include other commonly used tree growing softwares and routines (e.g rpart in R) or other classification approaches. However, the comparison is of two conceptually different ideas and seems a reasonable starting point.

Another criticism is that it is the analysis of just one data set on low birth weight. This data set was in fact stimulus for the work and, in terms of having relatively weak predictors, it is probably a fairly typical medical study. Other data sets where the ability to classify is greater could be studied, since it seems likely that sampling variability of partitions, and associated instability, may not be severe where there are combinations of predictors that clearly stand out.

Another criticism is whether any comparison of stability of methods that are intrinsically different is sensible, since each use different tuning parameters, which control to some extent the complexity of the partitions. For SPAN we have used default tuning settings and for CART our analyses were based on the marginally sub-optimal trees with at most 6 terminal nodes, the larger optimal trees being intrinsically more unstable.

One remedy for tree instability is to bootstrap re-sample at each node split and choose the split which is determined most often among the samples (Dannegger, 2000). This seems preferable to tree "bagging" (Breiman, 1996; Bauer and Kohavi, 1999) which is also suggested, but for which there is no single predictor tree. However, if logical Boolean structured classifiers are to be considered, using non-hierarchical alternatives to tree structured classification that are intrinsically more stable offers some promise. SPAN is considered in this paper, but there are others, for example, logic regression (Ruczinski *et al.*, 2003), rough sets (Deogun *et al.*, 1994) and hierarchical classes analysis (Leenen *et al.*, 2001).

Further, as we have also shown, after joining terminal nodes, tree classifiers often simplify and it can be argued, from principles of parsimony, that beginning with a class of simple irreducible partitions, is preferable to searching a class that is potentially reducible, as is the class of binary trees. Besides, even the simplest of partitions are likely to be overlooked by tree generation. For example, the simple overall best partition by SPAN for the low birth weight analysis (Table 3) is $(M)(W)(B)(S)$ which could only arise from a tree with a pathway to a single terminal $\overline{M}\ \overline{W}\ \overline{B}\ \overline{S}$ to represent $\overline{A}$, all other nodes combining to represent A. This structure was not generated by any of the trees.

When reported in the literature, tree partitions are not usually simplified, perhaps because the visual simplicity of a tree is more compelling than a Boolean expression, or because the Boolean reduction is tedious to achieve manually. In general, however, it would be helpful if simplification was a standard feature of tree analysis, since it would indicate redundant combinations and irregularities of the partition. In SPAN software there is a simple CART algorithm and the partitions of the trees that are grown are automatically simplified. That only

three of the 10 generated tree partitions in the Florida birth weight data turn out to be irregular may be a reflection of the relatively small size of the trees, since larger trees are invariably irregular. But it may also indicate that optimal partitions often are regular, and if this is so, a direct search of the space of regular partitions seems a preferable strategy.

References

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning* **36**, 105-139.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123-140.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, San Diego, CA.

Carmelli, D., Zhang, H. and Swan G. E. (1997). Obesity and 33-year follow-up for coronary heart disease and cancer mortality. *Epidemiology* **8**, 378-383.

CART. (2000). *Tree Structured Non-parametric Data Analysis.* Version 4.0. Salford Systems, San Diego, CA.

Dannegger, F. (2000). Tree stability diagnostics and some remedies for instability. *Statistics in Medicine* **19**, 475-491.

Dean, A. C., London E. D., Sugar, C. A., Kitchen C. M. R., Swanson A. N., Heinzerling K. G., Kalechstein, A. D. and Shoptaw, S. (2009). Predicting adherence to treatment for methamphetamine dependence from neuropsychological and drug use variables. *Drug and Alcohol Dependence* **105**, 48-55.

Deogun, J. S., Raghavan V. V. and Sever, H. (1994). Rough set based classification methods and extended decision tables. In *Proceedings of the International Workshop on Rough Sets and Soft Computing*, pp. 302-309. San Jose, CA.

Kitsantas, P., Hollander M. and Li, L. (2006). Using classification trees to assess low birth weight outcomes. *Artificial Intelligence in Medicine* **38**, 275-289.

Kitsantas, P., Moore T. W. and Sly D. F. (2007). Using classification trees to profile adolescent smoking behaviors. *Addictive Behaviors* **32**, 9-23.

Leenen, I., Leuven K. U. and Van Mechelen, I. (2001). An evaluation of two algorithms for hierarchical classes analysis. *Journal of Classification* **18**, 57-80.

Marshall, R. J. (1986). Partitioning methods for classification and decision making in medicine. *Statistics in Medicine* **5**, 517-526.

Marshall, R. J. (1995). A program to implement a search method identification of clinical subgroups. *Statistics in Medicine* **14**, 2645-2659.

Marshall, R. J. (2000). Generation of Boolean classification rules. In *Proceedings of Computational Statistics 2000*, pp. 355-360. Bethlehem, JG and van der Heijden, PGM (Eds.), Springer-Verlag.

Marshall, R. J. (2001). The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology* **54**, 603-609.

Marshall, R. J. (2006). Comparison of search partition analysis and other classification methods. *Statistics in Medicine* **25**, 3787-3797.

Nelson, L. M., Bloch, D. A., Longstreth, W. T. and Hong, S. (1998). Recursive partitioning for identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. *Journal of Clinical Epidemiology* **51**, 199-209.

Pawloski, L. R. and Kitsantas, P. (2008). Classification tree analysis of stunting in Malian adolescent girls. *American Journal of Human Biology* **20**, 285-291.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. L. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12**, 475-511.

Zhang, H. and Bracken M. B. (1995). Tree-based factor analysis of preterm delivery and small-for-gestational-age birth. *American Journal of Epidemiology* **141**, 70-78.

Zhang, H. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences.* Springer-Verlag, New York.

Zhang, H., Holford, T. and Bracken, M. B. (1996). A tree-based method of analysis for prospective studies. *Statistics in Medicine* **15**, 37-49.

Received April 12, 2011; accepted August 10, 2011.

Roger J. Marshall
Section of Epidemiology and Biostatistics
School of Population Health
Tamaki Innovation Campus
The University of Auckland, Private Bag 92019, Auckland, New Zealand
rj.marshall@auckland.ac.nz

Panagiota Kitsantas
Department of Health Administration and Policy
College of Health and Human Services
George Mason University
1J3 Northeast Module, 4400 University Blvd, Fairfax, Virginia 22030-4444 USA
pkitsant@gmu.edu