Proceedings of the 8th WSEAS International Conference on APPLIED MATHEMATICS, Tenerife, Spain, December 16-18, 2005 (pp144-149)

Discovering and understanding change using multivariate trees: the RECPAM approach

ANTONIO CIAMPI, ALINA DYACHENKO
Department of Epidemiology and Biostatistics
McGill University
1020 Pine Av. West, Montreal, QC, H3A 1A2
CANADA

http://www.medicine.mcgill.ca/epidemiology/ciampi/index.html

Abstract: - This paper was developed to understand a particular problem through data mining. A web-based questionnaire is used to allow visitors to evaluate a particular website and provide feedback to the owner. The questionnaire is proposed to the user over two time windows. We are interested in the following questions. Does the average evaluation change over time? Is this change uniform or does it vary across special visitor subgroups? If there are special subgroups, can we describe them? An answer to these questions is offered by tree-structured data mining, having as target the coefficient of the time variable of a multivariate predictor. The proposed approach is applicable to the general problem of detecting and understanding change when a flow of data is observed through several time windows.

Key-Words: -Prediction trees, Subgroup analysis, Recursive partition, Pruning, Amalgamation, Online questionnaire.

1 Introduction

Tree growing has been for quite some time a powerful tool for discovering structures in data. After pioneering early work in the sixties by a group of sociologists [1], trees were discovered and rediscovered by statisticians and computer scientists in the early eighties [2,3]. Since the mid-nineties, they have also become a basic tool in data mining [4]. For a recent review of work on trees, see [5]. Since the publication of [2, 3], only another monograph on trees has appeared [6].

The most popular tree growing algorithms aim to construct from data a tree-shaped rule for predicting a class variable or a continuous variable. On the other hand, there is a line of research, perhaps less widely known, that aims to construct tree-shaped predictors for some specified aspects of a complex phenomenon; these are usually represented by a parameter of a multivariate distribution or a stochastic process.

We have proposed a framework for developing tree growing algorithms adapted to such complex tasks, known as RECPAM, for RECursive Partition and Amalgamation [7-9]. In this paper we will present a particular application of this general approach, stemming from a practical problem. iPerceptions, a company which produces business intelligence for internet users, developed a few years ago a web based questionnaire for evaluating a website. Clients of this company are, typically, large companies offering goods or services through their own website. They wish to know whether visitors find their website attractive and useful, and they may want to use the information to make changes to their website. The questionnaire is offered to visitors. It can obtain in a very short time, from visitors who choose to respond, an evaluation of the website in the form of a five-dimensional profile, i.e. the values of five scales probing different aspects of the website.

Many iPerceptions client repeat the data collection in several time windows to follow the evolution of their visitors' preference over time. The website owner typically asks several broad questions, such as: a) What is the average profile? b) Does this profile vary substantially according to the characteristics of the visitor and the purpose of the visit? c) Does the average profile vary in time? d) Does the change in time, if any, depend on the characteristics of the visitors?

Question a) has an elementary answer: one simply defines a time-window and then calculates arithmetic means and standard deviations for the

five scales, taken over the population of users visiting the website in that time-window. This usually suffices to inform the website owner about the preference of visitors *over that particular time window*. A tree-structured predictor for multivariate response can answer question b). Taking a second time window and comparing the profiles obtained in the two windows provides answers to Question c). Question d) is the most challenging one, and the main focus of this paper is an attempt to answer it.

More generally, we will propose a tree-based data-mining solution to the problem of describing the time evolution of a multivariate response variable. The proposed approach aims to identify subgroups of subjects with distinct time evolutions and to describe them in terms of subject's characteristics. A theoretical treatment of the approach is beyond the scope of this paper and will appear elsewhere. Instead, we will describe here a practical solution to a common and interesting problem. Using the iPerceptions experience as a case study, we will highlight various features of the tree-growing methodology known as RECPAM and demonstrate its key role in understanding change.

2 Tree-growing for a multivariate response and the RECPAM approach

We will consider now the case of a client of iPerceptions, a hotel chain with a website describing their hotels and allowing various operations on line. As most clients, this one also wished to answer questions a) and b). Data were collected through the iPerceptions questionnaire described above, over a particular time window W_1. The five dimensions explored by the questionnaire are: Content, Interaction, Adoption, Motivation, and Navigation. The assessment of each dimension results in the value of a scale defined to be in the range [0,10]. The iPerceptions instrument also asks a few questions which provide information on the visitor.

Fig.1 answers question a): it shows the average profile of the five scales calculated from the data collected in W_1. Clearly, Content and Navigation are rated better than the remaining dimensions.

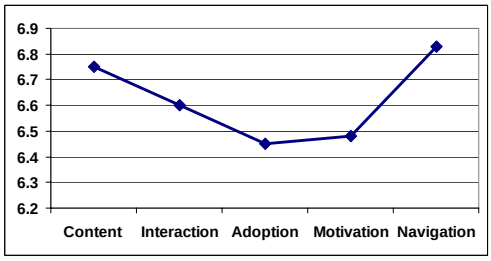

Fig.1 Average of 5 scales over sample in W_1

As an answer to question b), we present in Fig.2 a tree analysis of the same data obtained applying a RECPAM algorithm. The aim of the analysis is to understand in what way the characteristics of the visitors obtained from the questionnaire may influence their evaluation of the website. These characteristics are: Purpose of visit, Visitor group, Type of membership in a preferred guest program, Visit frequency, Total trips per year and Hotel name.

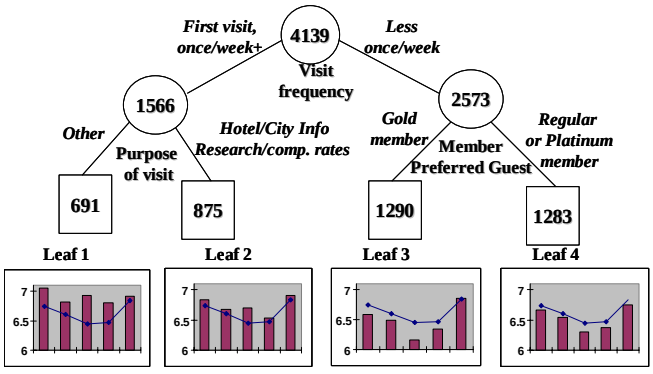

Fig.2 RECPAM Tree to predict profile
Bars indicate average scores over the tree leaf and the line represents the average profile over the whole sample

The interpretation of the figure is easy. For instance, it is clear that Navigation is seen as a strong point by all visitors group; that the group consisting of first time visitors and visitors who visit once a week or more often, for purpose other than information and rate comparison, seem to like the website considerably better than the average; and that, in contrast, golden members who visit less than once a week evaluate the website more severely than the average. Notice also that not all the predictors enter the tree: this is not due to the analyst's decision, but is the result of the algorithm's automated choices.

The process of tree construction for this case has been described elsewhere [10], but we will briefly outline it here to provide a non-technical explanation of the underlying methodology, which

is also common to the algorithm described in the next section. Assuming a multivariate normal distribution for the 5 scales, the goal of the tree construction is to predict the multivariate mean of this distribution, μ. The variance-covariance matrix V is of secondary importance here, and we do not wish to have it directly affect the tree construction; therefore we assume that it varies very finely across the sample: in RECPAM language, we treat it as a *virtual* parameter. The algorithm proceeds as follows: at each node of the tree, it chooses the variable and the split that contains the largest amount of *information* about μ. Information is defined as the likelihood ratio statistic of the hypothesis that μ varies across the split to the hypothesis that it does not—while V is allowed to vary across the split in the two hypotheses. Thus the search for optimal splits is entirely driven by μ and V only plays a subsidiary role. Notice that CART achieves the same goal in the one-dimensional case, but it does so by forcing the variance of the response variable to be constant throughout the sample, an assumption that is rarely justifiable. A repeated application of this search, with stopping rules involving only node size, leads to the construction of a large tree, just as in the CART algorithm. Next, again following CART, the large tree is pruned to reduce overfitting bias. Optionally, amalgamation of leaves from different parents may be performed, but no amalgamation was necessary in the examples treated in this paper, since leaves from different parents were recognized as quite distinct by the amalgamation algorithm in the examples considered.

To summarize, a RECPAM algorithm proceeds in three steps: 1) RECursive partition to obtain a large tree, 2) Pruning of the large tree, and, optionally, 3) Amalgamation. Unique to the RECPAM approach is the possibility of developing trees for the prediction of particular aspects of a complex phenomenon while correcting for secondary features. The example discussed above can be now extended to produce a tool for studying change.

3 RECPAM Tree-structured subgroup analysis as a tool for studying change

We can now return to the central problem of this paper: discovering and understanding change. Consider, as in our example, a multivariate response measured on a set of n subjects, each responding in one and only one of two time windows W_1 and W_2. It should be noted that the iPerception software does not keep track of the individual visitors. While it is possible that a subject be both in W_1 and W_2, this is not recorded, and so there is no way to model correlations. On the other hand it is reasonable to assume that if the time windows are sufficiently narrow, the number of subjects present in both of them is negligible; therefore our set up seems justified. Assuming multivariate normality for the response in both time windows, but with possibly different parameters, we can model the time change by a multivariate regression model with one binary variable. For the i-th subject we write:

$$Y_i = B_0 + I_{\{i\ in W_2\}} B_1 + \varepsilon_i \qquad (1)$$

where: i) Y_i is a row-vector of length p, the number of measurements taken on the *i*-th subject (p=5 in our example); ii) $I_{\{i\ in W_2\}}$ is an indicator variable taking value 0 for subjects in W_1 and 1 for subjects in W_2; iii) B_0 and B_1 are row vectors of parameters to be estimated from the data: in our example, they represent respectively the expected values of the 5 scales in W_1 and the difference between expected values at W_2 and W_1, i.e. the change occurring over time; and v) ε_i is a row vector representing the 'error term', assumed to have a multivariate normal distribution with expected value 0.

Equation (1) describes the situation in which subject characteristics do not affect the expected value of Y_i through the parameters B_0 and B_1. In this case, the estimation problem is trivial: B_0 is estimated as the sample mean of Y over the subjects in W_1 and B_1 as the difference between the sample mean of Y over the subjects in W_1 and the sample mean of Y over the subjects in W_2.

The 2-window model can be easily generalized to the case of K time windows: W_1, W_2,…,W_K. Indeed the model can be written exactly as in equation (1), but now the B_1 is a (k-1)×p matrices and $I_{\{i\ in\ W_k | k=2,..,K-1\}}$ denotes a row vector of K-1 matrix indicator variables, one for each window other than W_1:

$$Y_i = B_0 + I_{\{i\ in\ W_k | k=2,..,K-1\}} B_1 + \varepsilon_i$$

We can also model the effect of time explicitly. For instance, if we denote by t_k the mid-point of W_k, k = 1,…,K, then we can assume that change in the average profile is a linear or low-degree

polynomial function of time. Again equation (1) holds, but $I_{\{i\ in W_2\}}$ is replaced by a row vector of appropriate simple functions of the t_k's; this can be quite useful, especially if the saving in the number of parameter is substantial. Parameter estimation becomes slightly more complex, but the estimator can still be obtained in simple closed form. In what follows, however, we will continue to work explicitly with the 2-window model.

How can we discover change and explain it in terms of subject's characteristics? Clearly, change is detected if B_1 in the model of equation (1) is significantly different from 0. This, however, is only the starting point, since our goal is to discover whether or not the hypothesis of homogeneity of B_1 throughout the sample is supported by the data. The solution we propose is a tree-growing algorithm of the RECPAM family, i.e. an algorithm consisting of the three main steps briefly outlined in the previous section. Indeed, a RECPAM construction entirely and exclusively driven by B_1, would yield subgroups described in terms of the predictors, with homogeneous and distinct values of B_1. All other parameters would be 'virtual': they do not drive the tree construction, yet they are allowed to vary very finely across the sample in order to control bias resulting from omission of important parameters. Such a tree would contribute to the understanding of the observed change, since such a change could be related to the characteristics that define the subgroups.

We omit for reason of space the details of the implementation of the proposed algorithm. However, we wish to emphasize an important element of the construction. Recall that in the RECPAM approach to algorithm development, the key element to specify is the measure of information that a split contributes to the parameter driving the tree construction, in this case B_1 (leaf parameter). Our measure of information is defined as the likelihood ratio statistics (LRS) that compares the hypothesis that B_1 vary across the split (H_1) to the hypothesis that B_1 is the same in the two subpopulation defined by the split (H_0); in both hypotheses, B_0 and the variance-covariance matrix of Y are allowed to vary across the split, i.e. they are virtual parameters. The LRS is defined as twice the difference of two log-likelihoods, maximized under hypotheses H_0 and H_1 respectively. These hypotheses correspond to the following two models:

$$H_0: \begin{cases} E(Y_i) = B_0 + I_{\{i\ \text{in}\ W_2\}} B_1 + \\ \qquad + I_{\{i \text{ in left branch}\}} \Delta B_0 \\ Var(Y_i) = V + I_{\{i \text{ in left branch}\}} \Delta V \end{cases} \quad (2)$$

$$H_1: \begin{cases} E(Y_i) = B_0 + I_{\{i\ \text{in}\ W_2\}} B_1 + \\ \qquad + I_{\{i \text{ in left branch}\}} (\Delta B_0 + I_{\{i\ \text{in}\ W_2\}} \Delta B_1) \\ Var(Y_i) = V + I_{\{i \text{ in left branch}\}} \Delta V \end{cases} \quad (3)$$

where $I_{\{i \text{ in left branch}\}}$ is the indicator variable of the split, equal to 1 for subjects in the left leaf and zero for subjects in the right leaf, and ΔB_k, k =0,1, ΔV denote the difference of parameter values at the left and right leaf. It is important to notice that in model (2), both intercept vector and variance-covariance matrix are free to vary over the two leaves, while the slope parameter is assumed not to change from one leaf to another. By contrast, in model (3) all parameters vary over the leaves. It is this detail that assures that the algorithm is entirely driven by the slope parameter.

To the best of our knowledge, the approach presented here, in particular the treatment of B_0 and V as virtual parameters, is new. On the other hand, the idea of building a tree with regression equations at the leaves, which is also a RECPAM option [7], has been proposed earlier and by several authors [11-13]. However, these authors limit themselves to the prediction of a univariate response. More importantly, they ignore the distinction between leaf and virtual parameters. Therefore, their tree-growing algorithms are driven by both constant term and slope, which may result in an excessive number of splits of limited utility.

It is worth mentioning that the RECPAM approach has already been applied to the problem of predicting a slope in a regression equation (subgroup analysis). However, this has only been done explicitly for univariate and censored responses [7]. See also [14] for a detailed study of subgroup analysis for censored survival outcome.

4 Application to the Hotel chain data

We return now to the data collected by iPerceptions for the Hotel chain. The client, in response to the first data analysis, introduced substantial changes in the website. It should be stressed that the early analysis had not been accompanied by any specific advise on how to change the website, so that the

client was, and knew he was, entirely responsible for the changes. In fact, pleased with the first analysis, the client asked iPerceptions to perform a new data analysis with data collected a few weeks after the change. The average profile of the new data is presented in Fig. 4 and it clearly shows that visitors coming to the website after the change gave an even more severe evaluation than visitors before the change.

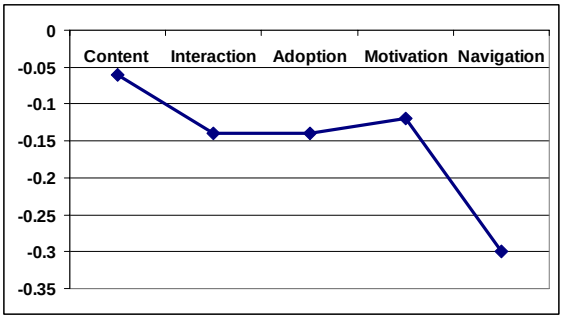

Fig.3 Average changes in the mean scores from W_1 to W_2

Obviously, the client was disappointed and asked iPerceptions to analyze the new data further to help understand the results. The main questions could be formulated as follows: are the disappointing results uniform across the new sample, or are there subgroups for which the change had had a positive impact? Is it possible to provide a simple description of such subgroups? The new analysis used the tree-growing approach outlined in the previous section. Thus the tree-construction was driven by the coefficient of the indicator of the new time-window. The results are shown in Fig.4.

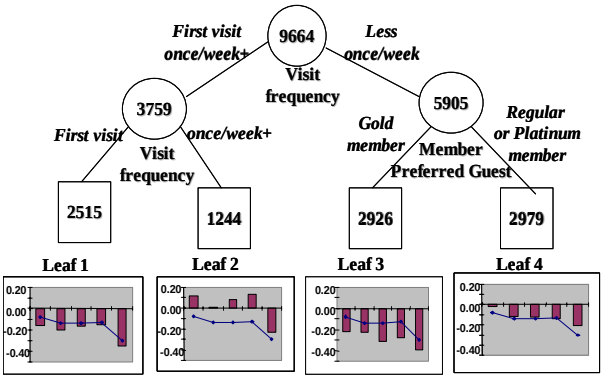

Fig.4 RECPAM Tree to predict change
Bars indicate average change in the 5 scores over the tree leaf and the line represents the average profile change over the whole sample

Interestingly, the tree identifies the subgroup of frequent visitors as the one for which a higher appreciation of the website was observed. In contrast, the golden members who visit less than once a week were those in which the largest drop in evaluation was observed; notice that this group is the same who had given the worst evaluation in the first analysis. Another important aspect emerging from the analysis of Fig.4 is that the evaluation of the Navigation dimension had considerably worsened for *all* subgroup of visitors. This suggests the following interpretation: the changes to the website, aimed at improving its attractiveness and usefulness, had actually resulted in a more difficult navigation; this increased difficulty may have caused a general irritation with the website for most visitors, except for the frequent ones, who were probably motivated enough by the new features to accept the increased complexity of navigation through the site. The client found the interpretation interesting and was pleased that a non-negligible portion of the potential market (frequent visitors are likely to be faithful clients of the hotel chain) had actually shown some appreciation for the new version of the website.

5 Conclusions

Since data mining became popular with decision makers in diverse areas, old and new tree-growing algorithms have been developed. RECPAM is a family of tree-growing algorithms similar to others, but with some distinguishing unique features. In this work we have illustrated through a real-life example some of these features. Indeed, RECPAM is model based and allows the user to focus on a particular aspect (parameter) of the model. The user selects the aspect of particular interest, and the algorithm construct a tree through a sequence of steps exclusively driven by this aspect. Other aspects (nuisance parameters) are not ignored but their role is an auxiliary one: they intervene only to control bias in the choices made at each step of the algorithm.

We have shown in particular how an algorithm of the RECPAM family permits to study time varying processes, identifying changes in time and providing suggestions for a first intuitive explanation of observed change. The unique features of the construction permit focusing on the parameters specifically describing change, while correcting for the impact of predictors on baseline values, variances and correlations. It should be emphasized that the algorithm itself does not

identify causation pathways, but simply offers a basis for discussing various hypotheses.

We have told the story of the application of RECPAM to a data-mining problem, just as it happened, except for changes in secondary details to preserve confidentiality. We feel that it shows quite clearly both the promises and limitations of RECPAM in a marketing context: RECPAM tree analysis seems to provide an insightful guidance in decision making, though it should be used with a good dose of prudence and without claims to infallibility.
As a direction for future research, we mention the development of appropriate RECPAM algorithms for studying change when a non-negligible number of subjects is present in several time windows.

References:

[1]. Morgan, J.N. and Sonquist, J.A., Problems in the analysis of survey data and a proposal, *Journal of the American Statistical association*, Vol.58, No.302, 1963, pp.415-434.

[2]. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Classification and Regression Trees*, Wadsworth: Belmont, CA, 1984.

[3]. Quinlan, J.R. Induction of Decision Trees, *Machine Learning*, Vol.1, No.1, 1986, pp.81-106.

[4]. Hand, D.J., Heikki, M. and Smyth, P., *Principles of Data Mining*, MIT press: Cambridge, Mass., 2001.

[5]. Ciampi, A. Prediction Trees, *Encyclopedia of Biopharmaceutical Statistics*, June 2005 update, Dekker Encyclopedias, 2005.

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

[7]. Ciampi, A., Generalized Regression Trees, *Computational Statistics and Data Analysis*, Vol.12, No.1, 1991, pp.57-78.

[8]. Ciampi, A. Constructing prediction trees from data: the RECPAM approach, *Proceedings of the Prague '91 Summer School of Computational Aspects of Model Choice*, Physica-Verlag, Heidelberg, 1992, pp.105-152.

[9]. Ciampi, A., Zighed, D.A., and Clech, J. Trees and induction graphs for multivariate response, *Principles of Data Mining and Knowledge Discovery*, (Zighed, D.A., Komorowski, J. and Zytkow, J. Eds.), Springer, 4th European Conference, PKDD, Lyon, France, 2000, pp.359-366.

[10]. Ciampi A., Arbres de prédiction pour variables multidimensionnelles, *Actes du IX-ème Congrès de la Société Francophone de Classification*, Toulouse, 2002, pp.25-33.

[11]. Chaudhuri, P., Huang, M.-C., Loh, W.-Y. and Yao, R., Piecewise-polynomial regression trees, *Statistica Sinica*, Vol.4, No.1, 1994, pp.143-167.

[12]. Chaudhuri, P., Lo, W.-D., Loh, W.-Y.and Yang, C.-C., Generalized regression trees. *Statistica Sinica*, Vol.5, No.2, 1995, pp.641-666.

[13]. Chipman, H., George, E. and McCulloch, R., Bayesian Treed Models, *Machine Learning*, Vol.48, No.1-3, 2002, pp.299-320.

[14]. Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S. and Boivin, J.F. "Tree-structured Subgroup Analysis for Censored Survival Data: Validation of Computationally Inexpensive Model Selection Criteria, *Statistics and Computing*, Vol.15, No.3, pp.231-239.