Using Regression Techniques for the Automated Selection of Radiosurgery Plans

Lisa McCrickard† Sven Koenig‡ Tim Fox§ Norberto Ezquerra‡

†GVU Center and
Bioengineering Program
Georgia Institute of Technology
Atlanta, GA USA
wenner@cc.gatech.edu

‡GVU Center and
College of Computing
Georgia Institute of Technology
Atlanta, GA USA
{skoenig,norberto}@cc.gatech.edu

§Department of
Radiation Oncology
Emory University
Atlanta, GA USA
tim@radonc.emory.org

ABSTRACT

The goal of our research is to automatically score radiosurgery plans. Radiosurgery is a technique for treating brain lesions with high dose radiation. When selecting a plan, the clinician must tradeoff between the complexity of executing the plan and the conformity in delivering the radiation. This decision depends on the treatment preferences of the clinician. Our research studies whether it is possible to learn to score radiosurgery plans in the same way the clinician does. We use regression to learn a preference function that maps plan properties such as complexity and conformity to the level of satisfaction that the plan holds for the clinician. The preference function is then used to predict the level of satisfaction that the clinician will have for unseen plans. The preference function makes it possible to either automatically select a radiosurgery plan for the clinician or make his task less time-consuming by sorting all plans according to their scores. We performed experiments with four different classes of preference functions and were able to identify the best plan with high reliability. We conclude that using regression to learn the treatment preferences of clinicians is a promising approach to automate or support the selection of radiosurgery plans in a clinician-specific manner.

Introduction

To develop decision-making tools a clinician will be willing to adopt, we need to take into account the treatment preferences of the clinician. These preferences, however, are something that clinicians cannot easily describe. They are based on personality, intuition, and clinical judgment. In this paper we therefore investigate whether we can learn them indirectly. We look at previous choices made by the clinician and see if we can use them to predict future decisions. We do this in the context of radiosurgery.

Radiosurgery is a specific form of radiation therapy. In it, high dose radiation is administered to a patient in a single treatment session. It has most commonly been used to treat diseases of the brain such as arteriovenous malformations, metastatic lesions, pituitary tumors, and acoustic neuromas [LFC00]. The document that describes exactly how the radiation is to be administered is called a *radiosurgery plan*. It details the prescription dose to be administered, the number and location of targets, and the beam orientations and intensities. Traditionally radiosurgery plans are created manually by a clinician interacting with planning software. However, this is a laborious and computation-intensive process. Because of this, automated plan generation has become an active area of research [DM97, MBB+98, LFC00]. Current methods can produce multiple candidate plans. One plan may differ from another by where dose is deposited or how many targets are used. It is necessary for a clinician to assess the clinical value of each plan and select the one which will be used to treat a patient. This decision is inherently clinician-specific, yet it requires the clinician to review prohibitively large amounts of graphical and numerical data.

Our goal is to automate the identification of the best plan or plans, in a clinician-specific manner. We want to learn a function that maps properties of a plan to a level of satisfaction that the plan holds for the clinician. This function serves as a preference model. We train the function on radiosurgery plans that the clinician has assigned satisfaction scores to. Learning is performed using regression, a mathematical technique for determining the parameters of a function that yield the least error between the function and a dataset. We then use this function to predict the level of satisfaction that the clinician will have for new, unseen plans. This function makes it possible to either automatically select a radiosurgery plan for the clinician or make his task less time-consuming by sorting all plans according to their scores. We developed the Automated Plan Evaluator to aid the clinician in the task of selecting a radiosurgery plan. The Automated Plan Evaluator uses the function to score the plans, and it then displays the top plan or plans for the clinician to assess.

Related work

Decision-analytic approaches have been used to automate the selection of radiation therapy plans [SO85, JK92, JKD+93]. Decision analysis is a method of making decisions based on the desirability of the outcomes, the probabilities of how known and unknown events will affect the outcomes, and the outcomes themselves [RN95]. Jain and Kahn [JK92] developed a multi-attribute utility model where the component attributes are the possible clinical complications of treatment. For each complication, the utility is calculated by combining the probability that the complication occurs with a weight representing the morbidity, or seriousness, of that compli-

cation. By assuming that each complication is independent of the others, they calculate the total utility of a plan by multiplying the effects of each complication:

$$\text{utility} = \prod_{i}^{\text{issues}} \left(1 - \text{probability}_i \cdot f(\text{prototypical\_weight}_i, \text{modifier}_i)\right)$$

Issues are the potential clinical complications such as a brain infarction or formation of a cataract in the lens of one eye. *Prototypical weight* is a measure of the consensual institutional conception of how important the issue is. The *modifier* is a weight that takes into consideration how the patient's case or the preference of the clinician differs from the institutional norms. *Probability* is the chance that a patient will suffer from that particular complication. A decision-analytic approach to scoring plans needs a probabilistic model of how radiation causes complications, and the model must be accurate. This approach tries to establish a score by reasoning about how radiation results in disease. In contrast, our approach based on a preference function attempts to strictly mimic the decision-making of a clinician. No attempts are made to reason about the correctness of the decision. Along similar lines, Willoughby, Starkschall et al. [WSJR96] have used artificial neural networks to score candidate plans. They train their network on a physician-assigned score and data on the volumes of tissue that receive different levels of radiation. In contrast, we look at different features of the plans and use regression techniques to score the plans to keep the amount of training data required to a minimum.

Description of Approach

Our method requires that we map properties of a plan to a satisfaction score. One question is what these properties are. We cannot choose too many of them because learning would then require too many training examples, and we want our Automated Plan Evaluator to quickly acquire the preferences of the clinician. If we select too few of them, we risk not conveying enough information about the plan. There are a number of figures of merit used in the field that summarize attributes of a plan, namely, *conformity*, *complexity*, *homogeneity*, and *coverage* [Rad96]. These figures of merit are promising candidates to be properties that map a plan to a satisfaction score. In this first study we use only two of them, *conformity* and *complexity*.

Conformity is an approximate measure of how well a plan delivers the prescribed level of dose to the tumor volume and nowhere else. Conformity, also called PITV (for prescription isodose over tumor volume) [Rad96], is defined as the volume of tissue receiving radiation at the dose prescribed or greater, divided by the volume of the tumor. The assumption is that the tumor volume is completely enclosed by the volume receiving high dose or that the volume receiving high dose is completely enclosed by the tumor volume. Volume is measured from the medical images of the patient. Images, referred to as slices, are acquired at various depths. The area of the region of interest is measured on each slice. This measure is converted to a volume by multiplying it by the slice separation distance. Achieving a conformity value of one is the ideal as neither subjecting healthy tissue to high-dose radiation nor neglecting to treat part of the tumor volume with the prescribed dose is desirable. *Complexity* is a measure of how expensive it is to execute a plan. In this research, it is defined as the number of targets included in the plan. A target is the point where radiation is focused. Each additional target incurs a high cost in how difficult it is to execute the plan. In weighing conformity and complexity, there is a clear tradeoff that the clinician needs to make. The closer the conformity is to one, the better the plan is considered to be. The closer the complexity is to one, the easier the plan is to execute. Unfortunately, to improve conformity one generally needs to increase complexity. Looking at this tradeoff allows us to study the decision-making of a clinician in an area clearly subject to clinician-specific treatment preferences.

At this point we have defined the plan properties that will be used to map a plan to the level of satisfaction that the clinician holds for that plan. A schematic overview of our process is shown in Figure 1. The upper tier of modules depicts the training process by which we acquire the clinician's preference function. The lower tier of modules depicts how the Automated Plan Evaluator is used to evaluate plans. The training process requires multiple training examples. Each training example is comprised of the plan figures of merit described here plus a score of the clinician's level of satisfaction with the plan. How we elicit the clinician's satisfaction level, and thus his treatment preferences, is described in the next section.

Eliciting the preferences of the clinician

The clinician is asked to both rank order and assign a numerical score to each plan according to, specifically, which plan he would most likely use to treat the patient. This forces the clinician to make the tradeoff between technical superiority of the plan and feasibility of executing the plan in the clinic. The clinician is asked to assign a score of 0 to 9 to each plan. A score of 0 indicates a poor plan, a plan the clinician would not likely use to treat the patient; a score of 9 indicates an excellent plan, a plan the clinician would be pleased to use to treat the patient. The extrema of the scale are fixed, but the rest of the scale is intentionally left undefined to allow the clinician to allocate individual meaning to the values. The clinician could realistically assign a different score to two plans with the same values of complexity and conformity because the plans differ in ways that are not summarized by the two figures of merit. For this investigation, we address this problem by restricting the clinician to knowledge of the two figures of merit and no other plan data. Having the clinician rank the plans in addition to assigning scores gave us a way to immediately validate the scoring. In some cases the scores did not reflect the order in which the plans were ranked, and the clinician was asked to reconsider his rankings and scores. By correcting these inconsistencies, the clinician was forced to better understand the specific criterion by which he was being asked to score the plans and to think deeply about what the scores meant to him. The output of this phase is n scored plans. Plan i in $1...n$ has complexity X_i, conformity Y_i, and clinician-assigned score $score_i$. These are the training examples that are used to learn the preference function.

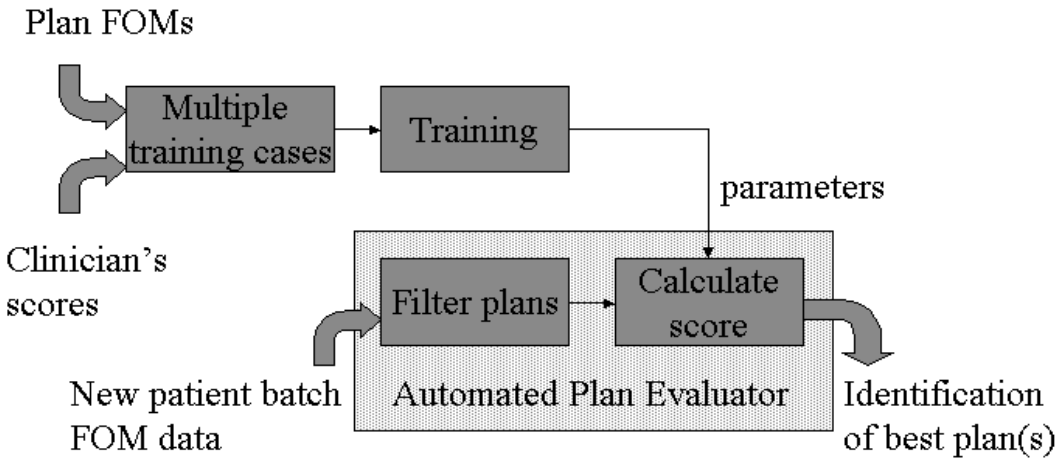

Figure 1: Plan figures of merit (FOMs) and the score assigned to the plan by a clinician are collected for several plans. The parameters of a clinician-specific preference function are learned by training a pre-selected class of functions on this data. This preference function becomes the core of the Automated Plan Evaluator. When given the figures of merit of an unseen group of plans, the Automated Plan Evaluator filters out the unsuitable plans and calculates a score for the remaining plans. It then identifies and displays the best plan or plans for review by the clinician based on this score.

Selecting the class of functions for regression

We want to learn a function from complexity and conformity to the clinician-assigned score from examples. In so doing, we want to generalize from known examples of the mapping to the mapping itself. The mapping can then be used to predict the scores of plans with a never-before-seen combination of complexity and conformity. We first need to select a class of functions that captures how the clinician makes the tradeoff between complexity and conformity. We then use regression to learn the parameters of the function. We thereby identify the single function that performs the mapping of the figures of merit to the score. Having more parameters in the function may allow us to fit the preferences of the clinician better. However, there is a tradeoff. More training examples are then needed to learn the parameters without over-fitting them, and training examples are expensive to obtain. Each one requires a significant investment of time by multiple clinicians. A physician must review the patient case and define exactly what volume needs to be treated and with what level of radiation. An assistant must format the patient and treatment data so that it can be read by an automated plan generator. Once the plans are generated, the figures of merit must be calculated. The last step is to have the clinician whose preferences are being studied score each plan. Because of this costly procedure, we would like to learn with a small number of training examples. Exactly what the class of functions must look like is not known *a priori*. We performed initial experiments to gain insight into what classes contain functions that can adequately mimic the decisions of the clinicians. In the experiments, which are described later, we tested four classes of functions having different numbers of parameters. In the training process depicted in Figure 1, the parameters of each class of functions were found using regression. The resulting functions were tested in the Automated Plan Evaluator. The single, best-performing function is subsequently used in the Automated Plan Evaluator to predict the scores of unseen plans.

Learning parameters by regression

In the previous section, we described how we select the class of functions to be trained on past cases. In this section, we describe how we use regression to specify its parameters. Regression is a mathematical technique for determining the parameters of a function that result in the smallest error between the function and a set of training examples. If we are given m training examples, where $score_i$ is the score given by the clinician to the ith training example, X_i is the complexity of the ith training example, Y_i is the conformity of the ith training example, and $score(X_i, Y_i)$ is the score of the ith training example predicted by the learned function, then the error is defined as

$$\sum_{i=1}^{m} (score(X_i, Y_i) - score_i)^2$$

Once the parameters are specified, we have identified the one function out of the class of functions that will be used to predict the score of unseen plans.

At this point, the Automated Plan Evaluator will take the complexity and conformity of an unseen plan and, using this function, map it to a score that predicts the level of satisfaction a clinician will feel for the plan. For the purpose of creating a clinically useful tool, we are interested in how plans score relative to one another. We compare against each other the plans that are candidates to treat a particular patient case. We call this group of plans a *batch*. In each batch of plans there can be plans that clearly would not be identified as the best. For example, they would have the same complexity as another plan in the batch, but they would have worse conformity. These plans are said to be dominated, and they are eliminated in a preliminary filtering step. A schematic drawing of how the Automated Plan Evaluator takes an unseen batch of plans, filters it, calculates the scores, and returns the best plan or plans for review by the clinician is shown in the lower tier of modules in Figure 1. The next

sections describe the experiments we conducted.

Experiments

The objective of our experiments was two-fold. We wished to evaluate how well our technique performed at identifying the best plan or plans from a group of candidates. We also wished to compare the performance of different classes of functions on this task. We had a clinician score a number of plans for different cases. We then had the Automated Plan Evaluator predict how the clinician would order the plans, and we compared the two. We did this for four different classes of functions.

Classes of preference functions

The four classes of functions we tested are listed below. In each, X is complexity, Y is conformity, and a_k is a parameter whose value regression estimates.

Log Model

$$score(X, Y) = a_0 + a_1 \log X + a_2 \log Y$$

Linear Model

$$score(X, Y) = a_0 + a_1 X + a_2 Y$$

Quadratic Model

$$\begin{aligned} score(X, Y) \quad = \quad & a_0 + a_1 X + a_2 X^2 + \\ & a_3 Y + a_4 Y^2 + a_5 XY \end{aligned}$$

Cubic Model

$$\begin{aligned} score(X, Y) \quad = \quad & a_0 + a_1 X + a_2 X^2 + a_3 Y + \\ & a_4 Y^2 + a_5 XY + a_6 X^3 + a_7 Y^3 + \\ & a_8 XY^2 + a_9 X^2 Y \end{aligned}$$

We chose the log function because similar functions appear in utility theory applications describing human decisions [KR93]. The other three are polynomial expressions of increasing complexity. We selected them in order to observe how increasing the number of parameters affects the accuracy of the Automated Plan Evaluator. All four classes are comprised of functions that are linear in the parameters making it simple to solve for them using regression.

Description of the data

Fifteen batches of plans were created for this study, a batch being a group of candidate plans created for a single patient case. Each batch contained the five plans that were not eliminated in the filtering step. Each of these have a unique number of targets, the number ranging from one to five. As the complexity of the plan is measured by the number of targets, each plan in the batch has a unique value of complexity.

The plans were generated by the mixed integer programming technique described in [LFC00]. A single clinician was asked to order and score all the plans. A total of five separate patient cases with unique anatomy and disease were used to generate the plans. Fifteen distinct batches of plans were acquired from this data, however, by varying other aspects of how the plans were generated. For the plan generation technique used here, a large number of candidate target points must first be identified. Only a fraction of them will ultimately be used in the resulting plan. The candidate targets are selected by sampling the treatment volume in a grid-like manner. The various batches differ in the number and spacing of candidate targets given to the plan generator. Additionally, two different algorithms, referred to as A and B, were used to computationally search for the best plan. Four batches were generated using a sampling grid spacing of 6 mm and search algorithm A. Five batches were generated using a sampling grid spacing of 6 mm and search algorithm B. Three batches each were generated using a sampling grid spacing of 4 mm and search algorithms A and B, respectively.

Validation method

Before the Automated Plan Evaluator can be used to predict plan preference, it must learn the preference function. Due to the small amount of data available to us, we could not partition the training examples into a group to use for training and another group to use for testing and still get meaningful results. Instead, we used leave-one-out cross-validation [Koh95]. In this technique the parameters are learned using the examples from all the batches, minus one. We use the single, reserved batch to test the performance of the Automated Plan Evaluator on unseen data and measure how well it does. The batch is subsequently returned to the set of batches used for training, and a different batch is removed to serve as the test batch. The process is repeated until each batch is withheld from the group of training examples one time. The error is the average across all iterations.

Metrics

There are two ways in which the Automated Plan Evaluator can be used. It could be used to select the treatment plan autonomously. In this case, the closeness in score between the plan the Automated Plan Evaluator identified as best and the plan the clinician defined as best is important. How suboptimal the treatment plan identified by the Automated Plan Evaluator is in the eyes of the clinician is measured by the regret. If *score(*$plan$*)* is the score assigned to plan $plan$ by the clinician, $plan_A$ is the plan identified as best by the Automated Plan Evaluator, and $plan_C$ is the plan defined as best by the clinician, then the regret is defined as

$$\text{regret} = |score(plan_A) - score(plan_C)|$$

The regret is zero if the plan identified as best by the Automated Plan Evaluator is given the same score as the plan the clinician ranked best.

However, it is unwise to completely remove the clinician from the decision process. A different way of using the Automated Plan Evaluator is to display all plans ranked in the same order as the clinician would rank them. This way, the Automated Plan Evaluator never excludes the best plan but still simplifies the work of the clinician by sorting them according to their scores. Thus, if the clinician has confidence

Batch	Log Model	Linear Model	Quadratic Model	Cubic Model
A	1	1	0	0
B	3	3	0	0
C	0	0	0	0
D	0	0	0	0
E	0	0	0	0
F	0	0	0	0
G	1	1	0	0
H	0	0	0	0
I	1	1	0	0
J	1	0	0	0
K	0	0	0	1
L	2	2	0	0
M	2	0	1	0
N	2	2	0	0
O	1	0	0	0
Max	3	3	1	1
Min	0	0	0	0
Mean	0.93	0.67	0.07	0.07

Table 1: Regret for the four functions fit using regression. Each batch is a group of five plans created for a particular patient case. The smaller the regret, the closer the match is between the score of the best plan of the five as identified by the preference function and the best plan of the five as defined by the clinician. A regret of 0 indicates the plans are scored equally.

Batch	Log Model	Linear Model	Quadratic Model	Cubic Model
A	0.60	0.60	0.90	0.90
B	0.30	0.40	1.00	1.00
C	1.00	1.00	1.00	1.00
D	1.00	1.00	1.00	0.90
E	1.00	1.00	0.90	0.80
F	1.00	1.00	1.00	0.90
G	0.60	0.60	0.70	0.70
H	1.00	0.90	0.40	0.30
I	0.60	0.70	0.90	0.90
J	0.90	1.00	0.90	0.90
K	0.40	0.70	0.70	0.80
L	-0.60	-0.30	0.70	0.90
M	0.30	0.90	0.90	1.00
N	0.10	0.20	1.00	0.90
O	-0.80	-0.60	0.50	0.90
Max	1.00	1.00	1.00	1.00
Min	-0.80	-0.60	0.40	0.30
Mean	0.49	0.61	0.83	0.85

Table 2: Rank correlation coefficients (RCC) for the four functions fit using regression. Each batch is a group of five plans created for a particular patient case. Values can range from 1 to -1. The closer the value to 1, the greater the agreement is between the rank ordering of the plans by the preference function and by the clinician.

in the capabilities of the Automated Plan Evaluator, he only needs to concentrate on the best-ranked plans. In this case, it is important that the Automated Plan Evaluator rank all plans in an order similar to that of the clinician, otherwise it would not be credible in the eyes of the clinician. This is evaluated by the rank correlation coefficient (RCC) [HH97]. We use it to measure the degree to which the plans have been ranked similarly by the Automated Plan Evaluator and the clinician. If $a = (a_1, a_2, ..., a_n)$ and $b = (b_1, b_2, ..., b_n)$ are two permutations of $1, 2, ..., n$ where a_i is the ranking of plan i by the Automated Plan Evaluator and b_i is the ranking of plan i by the clinician,

$$\mathrm{RCC}(a, b) = 1 - \frac{6 \sum_{i=1}^{n} (a_i - b_i)^2}{n^3 - n}$$

The RCC can range from 1 to -1. An RCC of 1 results when the two orderings are identical. An RCC of -1 results when one ordering is the reverse of the other. To evaluate how well the Automated Plan Evaluator performs, we look at both the regret and the rank correlation coefficient on several test examples.

Results

Table 1 shows the regret data for the four preference functions we tested. In all fifteen examples, the average regret was less than a single point. Two preference functions, log and linear, performed less well than the quadratic and cubic functions. In only one case each did the quadratic and cubic preference functions identify as best a plan that was not the best plan determined by the clinician. In those two cases, the identified plans were the clinician's second choice plans which were scored a single point lower than the first choice plans. Table 2 shows the rank correlation coefficients for the four preference functions we tested. The mean rank correlation coefficient was positive for all preference functions. The positive mean values indicate that the four preference functions are performing well above random in ordering the plans from best to worst. Again, we notice that the log and linear functions performed less well than the quadratic and cubic functions.

Discussion

According to both metrics, the quadratic and cubic preference functions performed better than the log and linear functions. The cubic preference function has ten parameters that are learned. The quadratic function has six. The two poorer-performing functions, log and linear, have three parameters each. More parameters allow us to fit the preference model of the clinician better. This is especially true if the preference model is complicated. However, more parameters require more training examples to learn them without over-fitting them. The risk is fitting the function to the noise of the data rather than the data as a whole. At this point, it is unclear whether the true preference model of the clinician is simple or not. It may be possible to reason that certain qualities characterize how the clinician makes his tradeoff decisions. These qualities could be captured in a cleverly chosen preference function without the ensuing need to increase the number of parameters to be fit. This might be possible by studying the preference functions themselves. Figure 2 shows the four preference functions that were determined from the training

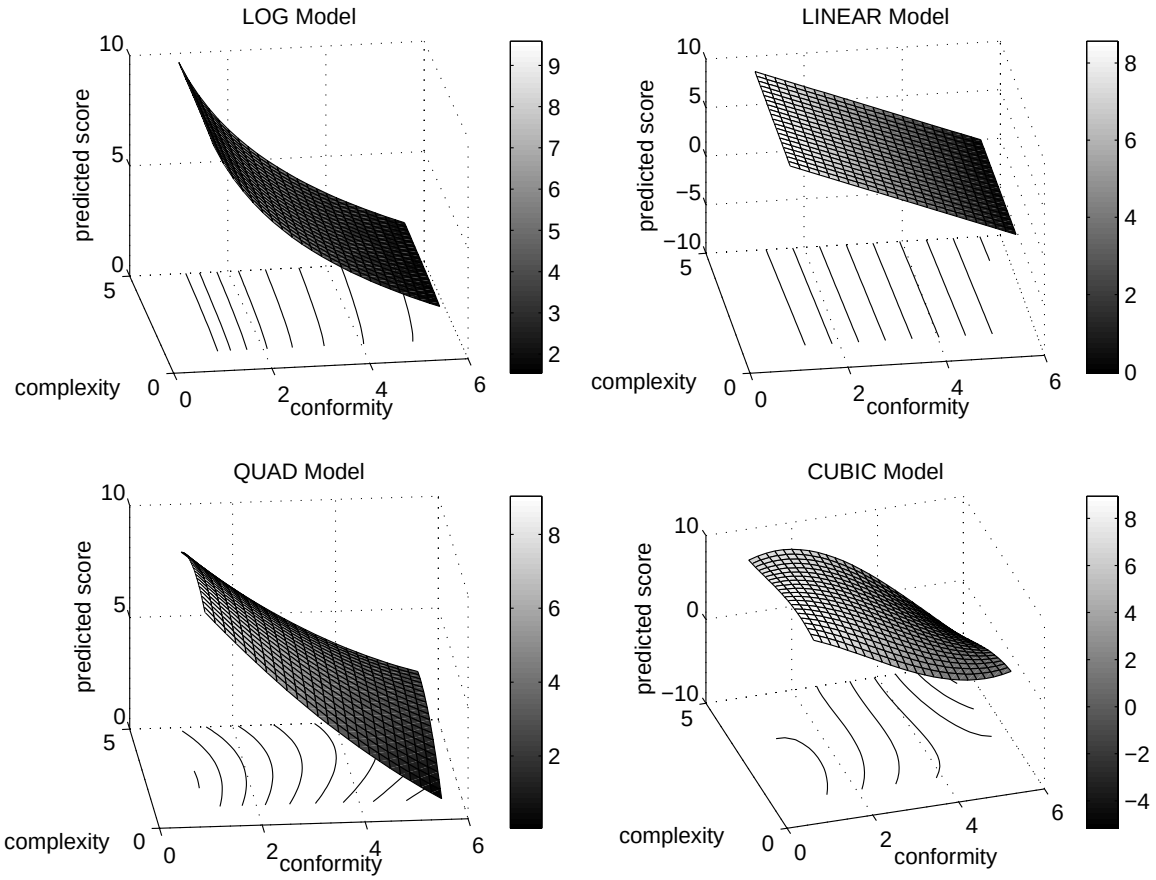

Figure 2: Plots of the four classes of preference functions that were fit using regression. Indifference curves, the curves indicating the combination of attributes the clinician is indifferent to, are projected onto the conformity-complexity plane. The models exhibiting greater curvature in their indifference curves, Quadratic (Quad) and Cubic, yielded better results in identifying the best plans in a clinician-specific manner.

data using regression. The predicted score as a function of complexity and conformity is shown on the vertical axis. It is also shown in the gray tones of the surfaces, lighter tones representing greater scores. Indifference curves are the contour lines of the surface projected onto the complexity-conformity plane, as seen in Figure 2. An indifference curve is a line that indicates all combinations of attributes that bring the same level of satisfaction to the subject, hence the subject is indifferent between these combinations. Examining Figure 2, one will notice that the two preference functions that produced better results, quadratic and cubic, have highly non-linear indifference curves. This indicates a highly non-linear tradeoff between the two attributes, complexity and conformity. The indifference curves of the quadratic function are concave over the region shown. The indifference curves of the cubic function are also concave except in the region of high complexity and high conformity where they are convex. Very few training examples were available for this range of the attribute space which makes it impossible to say whether the convexity of the indifference curves capture a true feature of the preference model. Instead, it could be an artifact of the small number of training examples we had in this range. The fact that the predicted scores in this region are quite inaccurate, in fact negative, supports the idea that the results in this region are spurious. If this is indeed the case, it may suggest that concavity truly characterizes the indifference curves of the clinician's preference function. Using results from utility theory [KR93], it may be possible to translate the curvature information into characteristics of an applicable preference function.

In our experiments, we tested two extreme cases in how we could use the Automated Plan Evaluator. Using regret, we looked at how well it can identify the best plan. Using the rank correlation coefficient, we looked at how well it can identify all the plans in order from best to worst. We got encouraging results under both metrics. In reality, however, we would most likely use the Automated Plan Evaluator in a manner that falls somewhere between these two extremes. We want to make the task of the clinician easier and less time-consuming, yet not remove him entirely from the decision process. We can do this by displaying, say, just the top three or five plans, in the order we predict that the clinician would rank them. This way the clinician has to evaluate only a small number of plans, yet it is very likely that the best plan is among them. When the clinician notices that the

Automated Plan Evaluator ranks the plans in the same way he does, the clinician will gain trust in it and begin to use it as a clinical tool.

Future work

This research was conducted under the premise that if a clinician's preference model can be quantified, it can be used to predict his selection of plans. The results gathered are promising. Utility theory may help us select a class of functions that will better capture the clinician's preferences, using as few parameters as possible. We also need to investigate whether the addition of other figures of merit or other plan attributes will allow us to make better predictions. Additionally, we wish to investigate how to learn the preference function using rankings of plans, rather than their scores. A weakness of the present technique of eliciting preference data as a score is that 1) the tradeoffs that must be made in the clinician's mind are not easily condensed into a single score, and 2) the scores may be context-dependent. That is, the clinician may consider the other plans in the batch when assigning a score. If, instead, rank data alone could be used to learn the parameters of the preference function, these difficulties could be avoided.

Conclusion

In this work, we examined how to automatically score radiosurgery plans while accounting for the individual preferences of the clinician. Our method is simple yet yields good results. We use regression to learn a preference function that maps properties of the radiosurgery plans to the level of satisfaction that the plans hold for the clinician. This preference function is then used to predict the level of satisfaction that the clinician will have for unseen plans. Our experiments showed that quadratic and cubic models were able to predict the best radiosurgery plans well and, to a lesser degree, rank-order all radiosurgery plans in a manner similar to that of the clinician. Our technique makes it possible to either automatically select a radiosurgery plan for the clinician or make his task less time-consuming by sorting all plans according to the level of satisfaction the clinician will likely assign to them. The work reported here is a first step in developing decision aids for clinicians, and it demonstrates the significant promise of using regression techniques for scoring plans and assisting clinicians in their decision making.

REFERENCES

DM97. S. Das and L. Marks. Selection of coplanar and noncoplanar beams using three-dimensional optimization based on maximum beam separation and minimized nontarget radiation. *International Journal of Radiation Oncology, Biology, and Physics*, 38(3):643–655, 1997.

HH97. Vu Ha and Peter Haddawy. Problem-focused incremental elicitation of multi-attribute utility models. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 215–222, August 1997.

JK92. Nilesh L. Jain and Micheal G. Kahn. Ranking radiotherapy treatment plans using decision-analytic and heuristic techniques. *Computers and Biomedical Research*, 25:374–383, 1992.

JKD+93. Nilesh L. Jain, Micheal G. Kahn, Robert E. Drzymala, Bahman E. Emami, and James A. Purdy. Objective evaluation of 3-D radiation treatment plans: a decision-analytic tool incorporating treatment preferences of radiation oncologists. *International Journal of Radiation Oncology, Biology, and Physics*, 26(2):321–333, 1993.

Koh95. Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.

KR93. Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, Cambridge, UK, 1993.

LFC00. Eva K. Lee, Tim Fox, and Ian Crocker. Optimization of radiosurgery treatment planning via mixed integer programming. *Medical Physics*, 27(5):995–1004, May 2000.

MBB+98. Sanford L. Meeks, John M. Buatti, F. Bova, W. Mendenhall, W. Friedman, and R. Zlotecki. Treatment planning optimization for linear accelerator radiosurgery. *International Journal of Radiation Oncology, Biology, and Physics*, 41(1):183–197, 1998.

Rad96. Radiation Therapy Oncology Group. *A Phase III trial comparing the use of radiosurgery followed by conventional radiotherapy with BCNU to conventional radiotherapy with BCNU for supratentorial glioblastoma multiforme*, 93-05 edition, 1996.

RN95. Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 1995.

SO85. Timothy E. Schultheiss and Colin G. Ortin. Models in radiotherapy: Definition of decision criteria. *Medical Physics*, 12(2):183–187, March/April 1985.

WSJR96. T.R. Willoughby, G. Starkschall, N.A. Janjan, and I.I. Rosen. Evaluation and scoring of radiotherapy treatment plans using an artificial neural network. *International Journal of Radiation Oncology, Biology, and Physics*, 34(4):923–930, 1996.