Board Game Recommendation

Sam Helmich

March 19, 2015

Abstract

You'll need to include an abstract here

1 Introduction

Board gaming, while being very popular among children, is a niche hobby beyond adolescence. Everyone may have grown up with Monopoly, but most people may have never heard of Citadels or Twilight Imperium. And as such, board gaming is a very difficult hobby to break into. "Gateway Games" such as Settlers of Catan or Ticket to Ride offer a gradual transition between the ultra-popular and mechanics-lite games like monopoly, to more rules-heavy games like Puerto Rico. For most people, the difficult part of the transition is actually finding new games to play. New games are often expensive for people who may be apprehensive about trying new games (a copy of Settlers of Catan goes for about $40 on Amazon). For this reason, a method to recommend games to people would be invaluable.

A few methods of varying usefulness already exist. Boardgamegeek.com, a popular site among frequent board gamers, has a recommendation system that allows users to find a particular game they like, and then from there it will recommend more games that a person may like. A pretty thorough analysis of it (and its issues) can be found at `http://boardgamegeek.com/wiki/page/Game_Recommendation_Algorithm`. The bigget issue is that it cannot be tailored to an individual. You, an individual user, may have different preferences than the userbase average, and it might be more useful to tailor recommendations to the individual.

Recommendations for the individual aren't easy to come by. There are a few automated systems out there (such as the boargamerecommender bot on reddit), but for the most part they are time intensive and often lack interactivity or any sort of information that would be useful to the user beyond a few single games that it might recommend.

In this paper, we seek to build a system the will recommend board games to users based on their own recommendations as well as the recommendations of others. This system will offer the user not only recommendations for a very large number of games, but it will also offer insights to the user about what sorts of games they like, not only scores for particular games.

2 Data Collection

For data, we turn to `boardgamegeek.com`, a website that holds information, ratings, and reviews for over 75,000 games. Users can create an account and then rate games they've played as well as keep an inventory of games they own. All of this information is accessable and scrapable, and will be the data that drives our analysis.

There are two types of particular information that we will be interested in: user ratings of the games and characteristics about the game (referred to as classifiers in the modeling section). The characteristics we're interested in fall into these categories as outlined in table 1.

Category	This covers the broad categories a game might fall into (ex.: Humor, Puzzle, Sports, etc.)
Family	Some games a naturally part of a "family" that share a name or common element (ex.: Animals:Bats, Ancient Wars Series, Hello Kitty)
Mechanic	This is what mechanisms the game uses (ex.: Dice Rolling, Card Drafting, Worker Placement)
Subdomain	More general than Category, but loosely bins games (ex.: Family Games, Strategy Games, etc.)

Table 1: Overview of the relevant characteristics of board games considered in this project.

Refer to table 2 somewhere in the text, add a label to the ratings table as well.

GAME_ID	CLASSIFIER	VALUE
1	yearpublished	1986
1	playingtime	240
1	boardgamemechanic	Area Control / Area Influence
1	boardgamemechanic	Auction/Bidding
1	boardgamefamily	Country: Germany
1	boardgamecategory	Dice
1	boardgamemechanic	Dice Rolling
1	boardgamecategory	Economic
1	boardgamemechanic	Hand Management
1	boardgamecategory	Negotiation

Table 2: Example lines of classifiers data, as seen in Database

The ratings for each game are on a scale from 1 to 10, with a score of 1 indicating a really bad game and a score of 10 being the best possible. Each of the ratings are paired with a unique user identification number as well as a game number. Combined with the classifiers data, we can provide very specific ratings for each individual user.

To keep the information as compact as possible, individual players and games are referred to by an ID number, which corresponds to a row in one of the key tables. These key tables contain information about the players and game that we don't necessarily need to include with each rating or classifier. This greatly reduces the size of the database needed to store and work with the information.

Not all of the data was scraped before creating the applicaiton. Data for 6,088 games was scraped and stored across the four tables. The reason for not scraping all of the data is twofold. First of all, for those 6,088 games there are over 3 million unique ratings. If we were to expand it to all 75,756 games known to BoardGameGeek, then our data set would be crippingly large for all but very high-powered machines. Instead, we implement a continuous data scraping procedure and update the recommender model dynamically: Each time a person uses the app, we make sure to check that all of their current information is correct. We look through known ratings and scrape their userpage on BoardGameGeek. Based on this, we revise old ratings and add new ones, as well as check to see that all of the games that this user has rated are included in the database!

	DATA_USERID	GAME_ID	RATING
1	734	1	10.00
2	820	1	10.00
3	1388	1	10.00
4	332	1	10.00
5	372	1	10.00
6	878	1	10.00
7	2587	1	10.00
8	628	1	10.00
9	5379	1	10.00
10	4899	1	10.00

Table 3: Ratings Data, as seen in Database

	data.userid	data.username	name
1	3	cwmassey	Craig Massey
2	4	greg	Greg Aleknevicus
3	5	gamemark	Mark Jackson
4	8	PBrennan	Patrick Brennan
5	9	JustDebbie	Deborah Pickett
6	17	gschloesser	Greg Schloesser
7	18	njsauer	Nick Sauer
8	21	mbialeck	Mike Bialecki
9	25	jaylorch	Jay Lorch
10	26	Chris Sjoholm	Chris Sjoholm

Table 4: The Player Key relates important info about the individual to the userid

3 Model

3.1 Model Selection

We base our implementation of a recommender system on two papers of on one the most famous recommender systems of all: Netflix. In 2006, Netflix offered a $1 Million prize for an improvement on their algorithm for predicting user rating on movies they had not seen. In 2008, two teams: BigChaos and BellKor (the eventual winner), published papers on their methodology. In the end, these teams ended up taking something of a shotgun approach, using a linear blend of many different models to come up with a single super-model.

XXX What is the approach you are taking compared to Netflix? the same? similar? what is different? - One or two sentences here.

There are a few things that make sense to model first. Certain games are more popular than other games, and certain people will naturally be more or less generous than others in their ratings. These two effects are very important to measure, as the first will allow us to provide estimates for people that have rated no games (and therefore we know nothing about), and the second will allow us to fine tune estimates for people who have rated many games. We will implement these in a similar manner to the way that BigChaos implemented the Global Effects model (**?**). XXX the next clause needs an end to be a sentence :) While they implemented 14 steps including Movie Effect, User Effect, and those variables crossed with variables like time since rating, average rating for the movie, production year, standard deviations, as well as others. XXX explain RMSE; make sure in the next paragraph to always point out what is yours and what is netflix. You know all those distinctions, but for any other reader it is not that obvious. From the RMSE of their probe set,

	game_id	name	year.pub	min.players	max.players	playing.time
1	1	Die Macher	1986	3	5	240
2	2	Dragonmaster	1981	3	4	30
3	3	Samurai	1998	2	4	45
4	4	Tal der Knige	1992	2	4	60
5	5	Acquire	1964	3	6	90
6	6	Mare Mediterraneum	1989	2	6	240
7	7	Cathedral	1978	2	2	20
8	8	Lords of Creation	1993	2	5	120
9	9	El Caballero	1998	2	4	90
10	10	Elfenland	1998	2	6	60

Table 5: The Game Key gives useful information about the game

the difference between using only user/movie effcts and using all 14 effects was only .02, a decrease of about 2%. Noting this negligible increase in accuracy and recognizing my lack of significant computing resources, we will only use user and game effects.

After we fit the general effects, we wish to model the remaining residuals, ideally in some way that gives us informative numeric summaries about the individual users in the process. In addition, we'd also like to compare a users ratings to other users, and hopefully use the behavior of simiar users to predict new ratings for someone. A tempting first step is to, for a particular game, find other people who have rated that game who have rated games that you've rated. Then, you could find people who rated games similarly to you, and use their ratings of the new game as an estimate. This often doesnt work, as sometimes the number of people who have rated the same games as you can be relatively low, if not 0 (show a graph for this part).

To get around the problem of low crossover in game ratings, we turn to the characteristics of the games themselves. This kills two birds with one stone. We'll create, for each user, a rating for each game mechanic and genre. Then, since there are so many mechanics and genres of game, a player only needs to rate a small, albeit diverse set of games before we can correlate that player's habits with other players, even if they've never played any of the same games! Then we can use those to filter rating results, as we'll know what board game mechanics that a particular player will enjoy.

We now have a very simple two step model that will allow us to predict ratings for new board games that a person may not have even played! Also, it allows us to rate games that nobody has played before: We can use the overall mean as a starting point, then for an individual player, we can adjust the rating based on their ratings of similar games from other genres.

3.2 Model Structure

The specifics of the model follow very closely to those outlined in the Global Effects and knn sections of the BigChaos paper [need bib]. Here, we cover the specifics of both.

Note for editing: I know this model is a mess at the moment, just getting the general ideas down on paper. this is not nearly as messy as you make it out to be, but you need to make sure to define every little piece of notation as you, because otherwise your reader will be lost.

3.2.1 Global Effects

The general idea behind global effects is to account for the natural biases present in particular games or players. For example, we might have two players who both play and enjoy the same game the same amount. However, one person might give this game a score of 10 while the other gives it a score of 9. Despite their overall enjoyment being the same, perhaps the latter player just tends to be a harsher rater in general. The global effect for this person would then be lower than that of the much more cheery fellow who gave the game a 10. Similarly, this effect can be present for games due to any number of reasons.

We'll estimate these effects one at a time using a heirarchical structure. After estimating the effect, we'll be left with a residual, which will be used as the response for the next level. The global effects formual is

$$r_{u,i} = \theta_u x_{u,i} + error \tag{1}$$

XXX what do you mean by first step? If you need iterations, use $r_{u,i}^{(n)}$ or something similar. If there is an iteration, make sure to give a convergence criterion. XXX: doesn't this error need to be indexed as well? could we call the error ϵ and make away with the text in the formula? XXX: generally: no future tense in the write up. The CC is in the here and now :)

Where for the first step, $r_{u,i}$ ~~will be~~ is the rating from user u for game i and $x_{u,i}$ will be an indicator variable that is 1 if player u has rated game i and 0 otherwise. In the second step $r_{u,i}$ will be the residual from the first portion, the subscript on θ becomes i and $x_{u,i}$ will be an indicator variable that is 1 if the rating is for game i and 0 otherwise.

XXX it is θ_u, not θ. Give a formula for what you are doing with the θ_u. XXX Why do you call α a fitting paramter and not a shrinking parameter?

To get an estimate of θ at either step, we'll use the average of the nonzero values which are shrunk by the fitting parameter α to avoid overfitting (motivation for using α in this manner can be found in Bell and Koren (2007)). The sum is over either all ratings for user u or all ratings for game i, depending on whether you're estimating θ_u or θ_i. Either way, the structure is the same.

$$\hat{\theta_u} = \frac{\sum r_{u,i}}{n_u + \alpha} \tag{2}$$

Noting that this can produce esimates that are above 10 and below 1, ratings that fall outside of $[1, 10]$ are truncated to the closes interval endpoint.

3.2.2 Classifier Aggregation

Independent of the global effects, we need to give each user a score in each classification category. We will use 4 types of categorical variables, which will yield (FIND OUT HOW MANY CLASSIFIER VARIABLES). To do this, we can either use the ratings themselves, or the residuals from the global effects process. Let $c_{u,j}$ be the categorical rating for user u in category j. Additionally, for each game i, there exists a set of classifiers C_i that describe the aspects of the game in things such as mechanics, game family, categories, and subdomains. Then for each user, we take all of the categorical values for each user and normalize them, so they should all have mean 0 and standard deviation of 1 for each user. Then, within each category, the average rating is calculated. Now, if a user hasn't rated anything in this category we can assign them a score of 0 in that category (what should amount to an average rating). These serve dual purposes: Firstly, we can assess what sorts of games a user rates higher than average most of the time, and game they rate lower than average most of the time. This will allow the algorithm to not only recommend specific games that have a high predicted rating, but will also allow for the user to find categories or mechanics that they enjoy as well. The other purpose is that we now have a larger dataset to compare people than we might have had before, as we are guarantedd to have (FIND OUT HOW MANY CLASSIFIER VARIABLES) for every user.

3.2.3 k-nearest-neighbors

After we aggregate on the classifiers, we can then find correlations between all of the users. We will then take the k nearest neighbors, where the distance is given by $\rho_{i,k}$, the correlation between the i and kth users. In this algorithm, k will be chosen by a crossvalidation procedure that will be done on a user-by-user basis, as determining a global value for k would be far too computationally expensive. After the optimal k is found, then for each game the k nearest neighbors' (who have rated the game) rating's are averaged using a weighted mean, with the weights being equal to $\frac{2}{1+\rho_{i,k}}$.

3.2.4 Back to Rating

After the k-nearest-neighbors step, we need to "rewind" from the global effects fitting process. For each game, we'll need to add on the specific user effect, the effect for each game, and the overall game mean. This will then yield a rating for each game for each player!

3.3 Parameter Optimization & Selection

There are three parameters that the model relies on: The two α values that balance out biasedness in the global effects section, and the k in the k-nearest neighbors portion of the algorithm. The α's will be done globally (that is, every individual will use the same values for α_1 and α_2), while the values for k will be done locally (each user will have their own value of k).

3.3.1 Optimizing α

To optimize the α values, we'll take all of the ratings for all players and subtract off the mean of all of the ratings (this is the first step in the globalized residual process). This leaves us with three variables: The userid, the gameid and the residual. We'll randomly split all of the data into 10 equally sized buckets. We'll use each of these buckets as a test group, where we put the other buckets together, fit a model using a canidate value for α, and then fit the values in the test group. From each test group, we'll then find the mean square error, and then average those together from all 10 groups. Taking the square root of that average, we can get the RMSE for each canidate value for α. Finally, we'll proceed using the canidate value for α with the lowest RMSE. A graph of this process can be seen in figure 1, with the domain truncated for easier visibility.

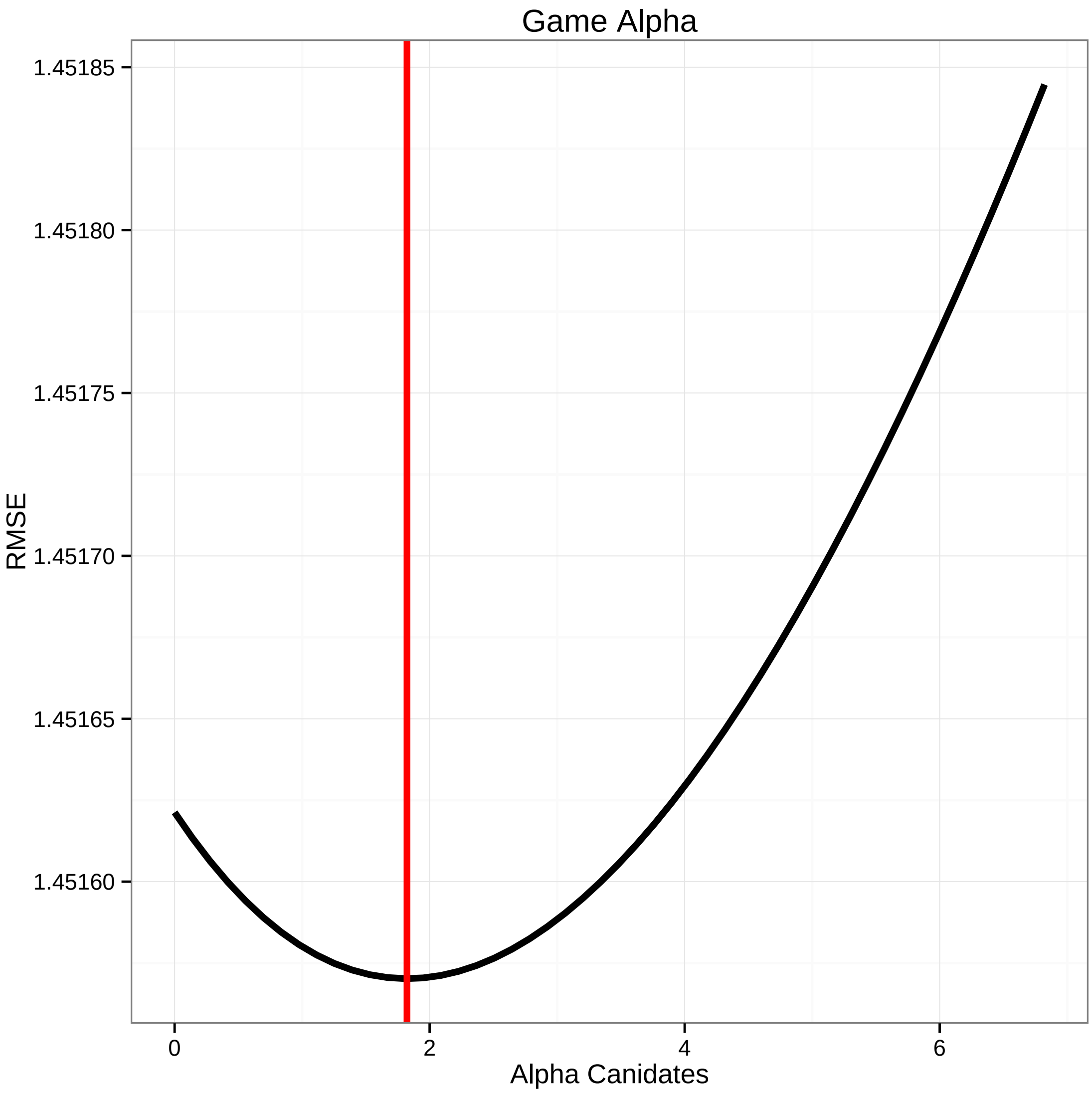
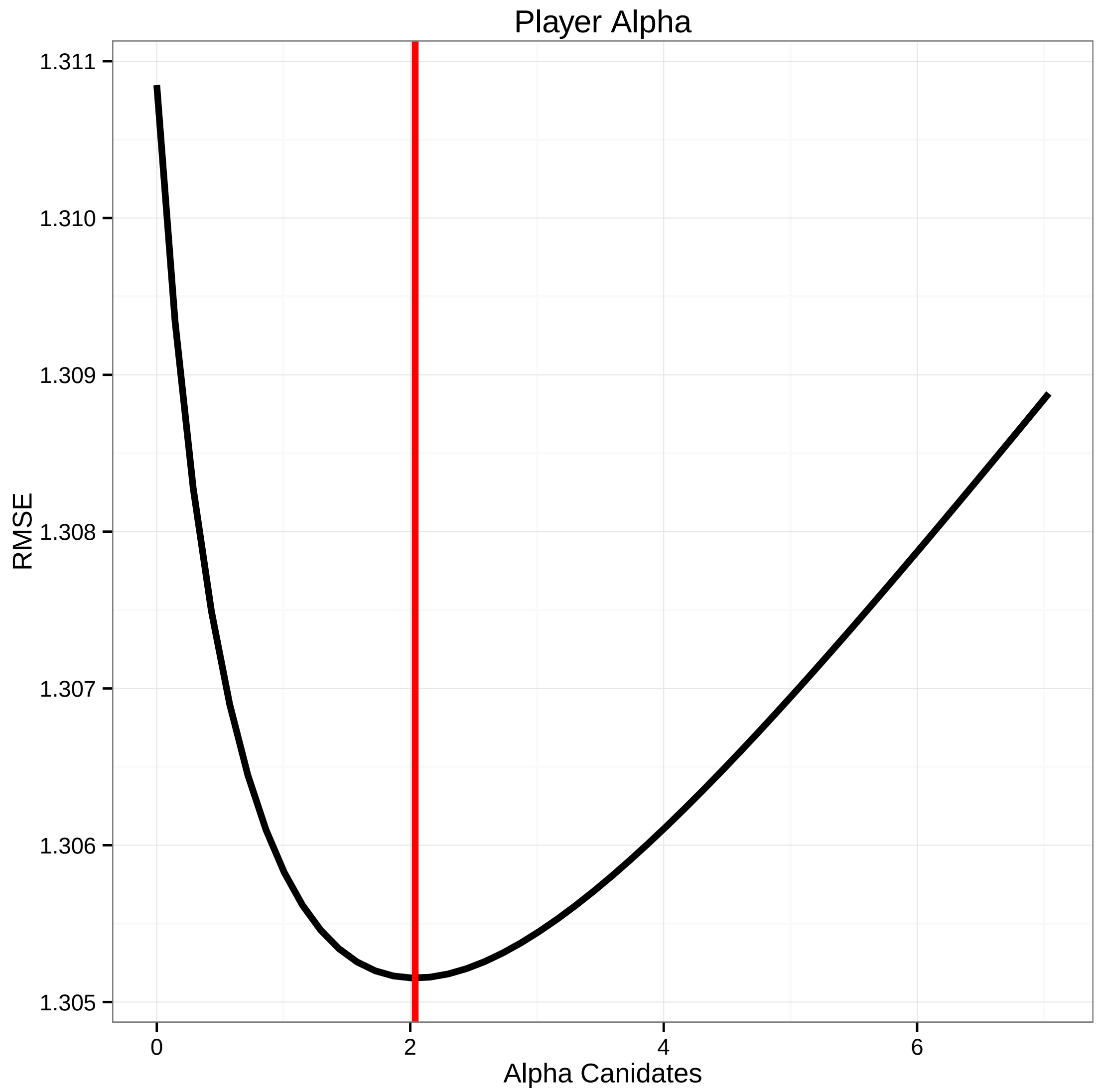
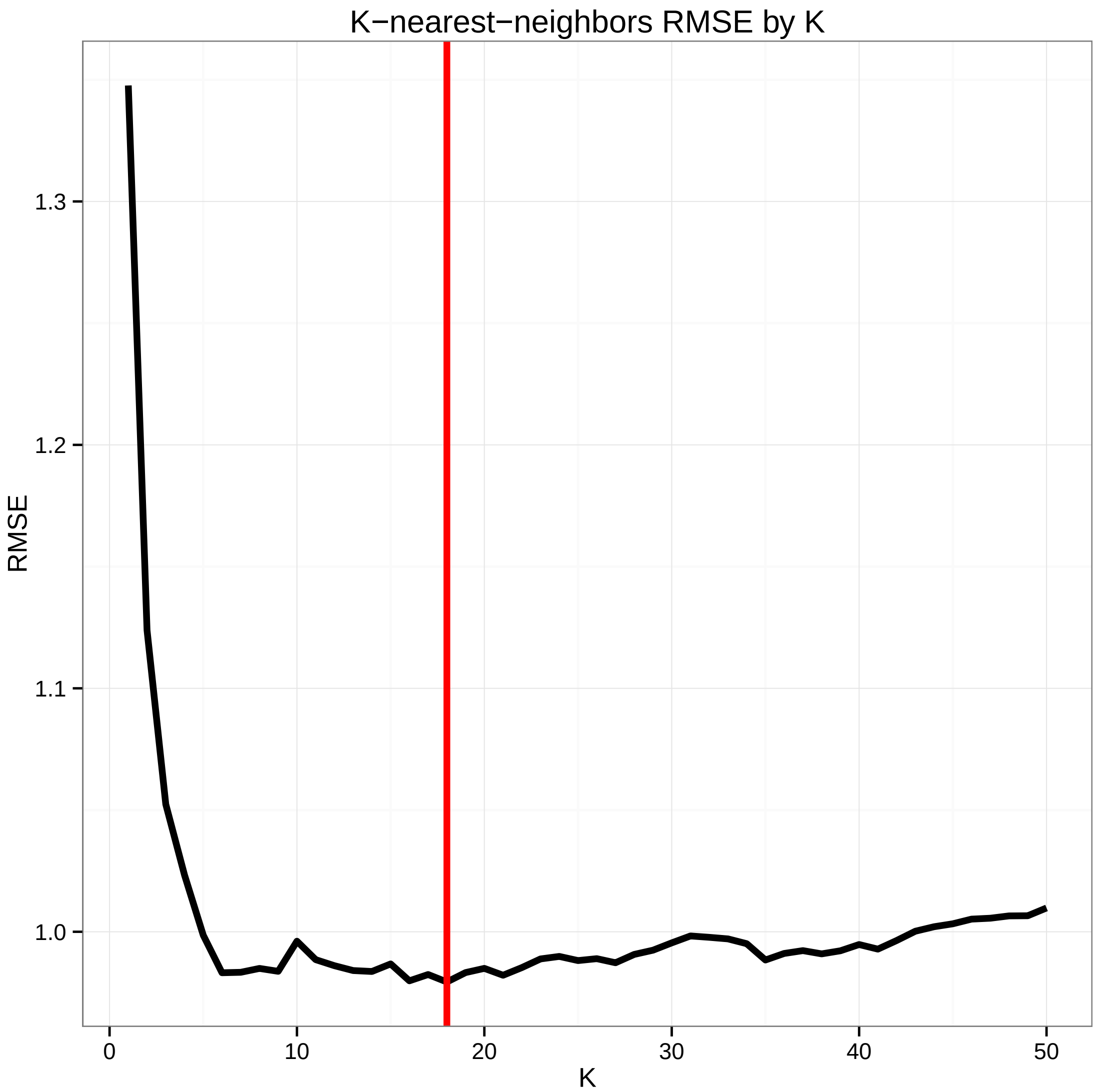


Figure 1: You need a caption here - what is shown?


Player Alpha
1.311
1.310
1.309
1.308
1.307
1.306
1.305
RMSE
0
2
4
6
Alpha Canidates



While the approaches for finding the optimal values for these parameters may be slightly different, the general implementation will be the same. We'll first begin by partitioning a subset of the data into 10 different sections (10 'folds'), and then choosing an appropriate parameter space for the parameter to be optimized. Ffor the alpha's, we'll choose values between 0 and 1000 (double check this), evenly spaced, and for k we'll choose a positive integer. We'll then proceed fold by fold, using the other 9 folds to fit a model, and then using the data in current fold to check how the model works. Using root mean square error, we can then rank each of the possible parameter values in the parameter space, and the value with the lowest rmse will be the 'optimal' value for that parameter.

4 Technical Implementation

** SCREENSHOTS NEEDED **

This algorithm is implemented in R using shiny as a vehhicle for a simple, easy to use user interface. Behind the scenes, the data is stored in a data base with six tables: A player key table, which relates the players information to a unique player userid; A game key table, which relates game information to a unique game id; A ratings table which stores the ratings (a score from 1-10) along with the userid of the person who rated it and the game id which the rating is for; the classifiers table, relates the game id to characteristics about the game that will be useful in the classifier aggregation and k-nearest-neighbors section; the predicted ratings table, which will store predicted ratigs if ratings have already been fit for a particular user; and the aggregated classifier ratings for each user. This data storage will allow for ease of implementation: we only really need the ratings table and the classifier table to do all the work, and the key tables are really only needed to display the inforation to the user in a format which they will understand.

In an implementation of this, the application will begin by having the user enter the username they've registered on `boardgamegeek.com`. Instead of logging the users ratings directly through the application, we will instead leverage the existing infrastructure of boardgame geek. After checking that its a valid username, the application will scour boardgamegeek to see if the database's information is complete: that is, do we have this user's data in the player key? Have they rated games that aren't in our database? Have they rerated games that are in our database? Have they rated games that aren't in our database? .If so, we also need to bring in other users ratings for any games that we may not have had in our database (since a single rating for a game isn't very useful and will greatly skew the results for that game). After this initial step and database update, we can see if they already have saved predicted ratings in the system, and if not, we'll proceed to model fitting.

We first apply the global effects strategy as descibed in the modling section. The optimization of the α parameters will have occred sometime before this and the algoritm will use use those stored values. A weeky or monthly automatic updaing procedure is used to keep these values up to date as more data comes into the system. From the global effects, we then aggregate those based on the classifier aggregation procedure as described above. These are then stored in the database for future use by the system. After that, before we move onto the k-nearest neighbors implementation, we must first find an optimal value for k using the 10-fold cross validation as described above.

Finally, after the k-fold crossvalidation, we're ready to run the other games through the algorithm to get a predicted score. The games that the user has already rated will then be replaced with those ratings, and the full list will be outputted for the user.

Since we saved the aggregated classifier ratings for the individual user, we can suggest ways for the user to subset this data. By providing the graph below to the player (GET THE GRAPH WORKING), they can see for themselves their not-necessarily obvious preferences in game family or mechanic. They can then search by their preferred mechanic or theme, and get ratings for games that fit that description. Or the user can simply sort by highest rated.

5 Conclusion

References

Bell, R. M. and Koren, Y. (2007), "Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights," .