Complexity in board games

Federico Fiorio
computer science department at Università degli Studi di Milano
Milan, Italy

I. INTRODUCTION

Defining board games complexity it's a very difficult task. In this paper some of the best features for determining the best possible predictions for game complexity were analyzed and chosen accordingly.
The BoardGamesGeek (BGG) database and APIs were fundamental to extract some important features not presented either in the definition of game complexity introduced by BGG, nor could be extracted from games rulebooks.
In the current literature, not much work has been done in the studying of board games, even though these games can be used as a tool in entertainment and research in fields like artificial intelligence and military.
In the paper "A Data Driven Review of Board Game Design and Interactions of Their Mechanics" [1] some work has been done towards the use of game complexity but for addressing design purposes.
In the paper "Yasseri, T., Kornai, A., Kertész, J. (2012). A practical approach to language complexity: a Wikipedia case study. PloS one, 7(11), e48386" [2] board games are not the case study but some important analysis has been done in order to compare a complex language and a simple language.
In this paper the features extracted from BGG database and games rulebooks were used for the prediction of game complexity.
These features are chosen because they lead to the least mean squared error (MSE) between the predictions of games weights and the target weights extracted from BGG database. The prediction scores are computed using a gradient boosting model. Finally, some suggestions are given in order to perform further studies on game learnability which is easier to compute with respect to game complexity and with a higher impact on real world applications.

II. RESEARCH QUESTION AND METHODOLOGY

In this paper the main aim is to define the methodology for finding the best group of features that can be extracted from the games rulebooks and the BGG data. These features can be used for the prediction of complexity of unseen/new board games using artificial intelligence approaches.
The target complexity which is used in the following work is called weight and it is based on users reviews.
The weight, in other words, is the perception of complexity of the users that played the game, some of them might be biased because it might happen that not a large quantity of users rated the game.
To avoid bias issues, in this paper, the vast majority of games used for the analysis were in the top 100 of BGG rated games or well-known, with some exceptions.
There are mainly two formal definitions of complexity and they are: 1: A whole made up of complicated or interrelated parts; 2: Group of obviously related units of which the degree and nature of the relationship is imperfectly known. BGG tries to give a definition of game weight, but it cannot be defined at BGG because this website wants to guarantee that different people have different ideas of what it means, thus BGG only gives factors that affect game weights. These factors are:

- Amount of rules;
- Gameplay length;
- Amount of luck;
- Technical skill required (math, planning, reading, etc.);
- Amount of choices available;
- Amount of bookkeeping;
- Level of difficulty (!), this feature is redundant.

While the majority of features have been approximated or computed, the amount of choices available is too hard to be generalized for a big quantity of board games.
It can be for sure computed for a single game given rules and map settings, but creating a method that is able to generalize the understanding of these rules and components so that, after this first phase it is possible to actually compute the amount of choices available for each player requires too many details to capture and a large quantity of data, that in some part can be retrieved from BGG, but gathering the rulebooks is external from BGG and requires other tools for a good automatization, also some rulebooks of some games cannot be found so easily, thus also the automatization search is not an easy task.
For these reasons this feature was omitted and thus some games that heavily base their weight score on the amount of choices/possibilities that a player have during the gameplay (Go, Chess ...) might not be well approximated. The BGG factors that influence a game's weight are a good starting point, but they are not enough.
During the analyses it was clear that extracting readability indices from the rulebooks could give a good approximation of the technical skill required for a person to play the game.
These indices are based on words/sentences length and they are used to understand what level of scholarization is needed to understand the given text. Since not all the readability

indices are computed in the same way, they also give different results, six different types of indices have been extracted from rulebooks, but actually, only two were chosen as the best.
Some readability indices are extracted on a random sample of the rulebook, for example on some random sentences, this approach can lead to different final best features, especially because the sample of board games used is not vast. The best features could also change if more samples are added, however the aim of this paper is not to give the best features ever computed for a board game to predict its weight, but it wants to give the methodology used for finding them. The methodology should guarantee that with a large enough sample, the randomity of some indices will not be an issue and also adding more samples to the already big dataset will not change the overall results.

The features extraction was achieved using two different sources, the first source was BGG dataset.
From BGG via the BGG's APIs the xml files of the games were extracted, they represent a semi-structured file with much information that could be used for the purpose of the work.
The main features taken into consideration from this source are: the average suggested playing age, the playtime, and the text dependency while playing the game. The average suggested playing age is not the average age of players, but the average age required to play the game. It should be clear that there is an abyssal difference between the average playing age and the suggested required age for playing the game, while the first one does not really give much information about the complexity of the game, the second one could clearly be an advice of how hard the game is, to play and to understand.
While there might be some outliers, for example, extremely easy games but with a high suggested playing age due to adult content, these cases are not the rule, and they are in a vast minority compared to the others.

The play time of a board game is a very powerful hint on the game complexity. It is true that also here there could be some outliers, for example Monopoly is not a difficult game to understand, the weight is quite low, but it could have a very long play time, however, board games tend to have a long gameplay only if there a lot of mechanics/choices to take into consideration or there are many phases with different rules (for example the fury of Dracula).
The intuition is simple, very few people are willing to play a three hours game if the gameplay is repetitive, in order to have a good design and a good gameplay the users must be kept attracted in every phase of the game, they must be willing to end the game, and not be bored when they finished it, thus it is in the interest of the game production industry to keep a short gameplay if the game does not have many mechanics. On the other hand, the game will have a bigger and more complex gameplay if it has a larger variety of scenarios to explore, mechanics to use or alternatives to choose.

The last feature extracted from the BGG APIs is called text dependency or language dependency. This feature goes on a scale from 1 to 5 where 1 means "No necessary in-game text" and 5 means "Unplayable in another language".
The in-game text does not refer to the rulebooks, the rules must be known when playing the game, the dependency that this feature is addressing is based on the content needed to play the game, for example: cards, instructions and so on.
Simple games with not many rules/mechanics and thus complexity, usually have a small value in text-dependency, you could easily memorize some specific cards or no need to memorize them at all to actually play the game.
The harder the game is, the easier is the possibility to have cards with longer text, or specific subplots with different scenarios, these events can occur during the gameplay and require the language knowledge to at least understand what is happening during the game.

The following features were extracted from the rulebooks, that is, the second source for gathering informations used in this paper.
A total of six readability features were extracted:

- Flesch reading ease;
- Gunning Fog formula;
- SMOG formula;
- Dale Chall readability score;
- Powers-Sumner-Kearl Readability Formula;
- FORCAST readability formula.

The Flesch reading ease is computed as:

$$ReadingEaseScore = 206.835 - (1.015 * ASL) - (84.6 * ASW)$$

Where:

- ASL = average sentence length (number of words divided by number of sentences);
- ASW = average word length in syllables (number of syllables divided by number of words)

Scores between 90.0 and 100.0 are considered easily understandable by an average 5th grader;
Scores between 60.0 and 70.0 are considered easily understood by 8th and 9th graders;
Scores between 0.0 and 30.0 are considered easily understood by college graduates.

The Gunning Fog formula is computed as:

$$0.4[(words/sentences) + 100 * (complexwords/words)]$$

Where:
complex words are identified as words with syllables equal or higher than 3. The range of values goes between 6 and 17 where 6 indicates a sixth grade level and 17 indicates a college graduate level.

The SMOG formula is computed as:

$$3 + (complexwords^{0.5})$$

Where:
complex words are words with 3 or more syllables taken from a specific sample of 30 sentences.
The sample is computed at random so it is possible to have different values for the same item.
The values range from 0 to 211+.
0 is the level of a 4th grade and 211+ is the level of a professional.

The Dale Chall readability score is obtained as:

$$0.1579 * (PDW) + 0.0496 * (ASL) + 3.6365$$

Where:

- PDW = Percentage of difficult words;
- ASL = Average sentence length;
- The difficult words are the ones that are not contained in the Dale Chall list.

This feature have values from 0 to higher than 10 where below 4 indicates a 4-th grade student or lower and 9.9 indicates a college student.

Powers-Sumner-Kearl Readability Formula is computed as:

$$RA = 0.0778 * (ASL) + 0.0455 * (NS) + 2.7971$$

Where:

- RA = Reading Age;
- ASL = Average Sentence Length;
- NS = Number of Syllables;

The formula is computed on a simple passage of 100 words, this sample is extracted at random in the rulebook, thus it is not guaranteed that the sample start and end will make sense.

the FORCAST readability formula is computed as:

$$GradeLevel = 20 - (N/10)$$

Where:

- N = number of single-syllable words in a 150-word sample;
- The 150 words are extracted from a portion of the text and they are not chosen singularly at random.

The next three computed features are:

- number of rules;
- luck;
- bookkeeping.

The number of rules were approximated with the number of sentences of the overall rulebook, it might lead to some outliers, because some rulebooks might contain many examples that repeat the already presented rules but do not insert new complexity; However, as shown in the following image, there is some correlation with the complexity of the game and the number of sentences.

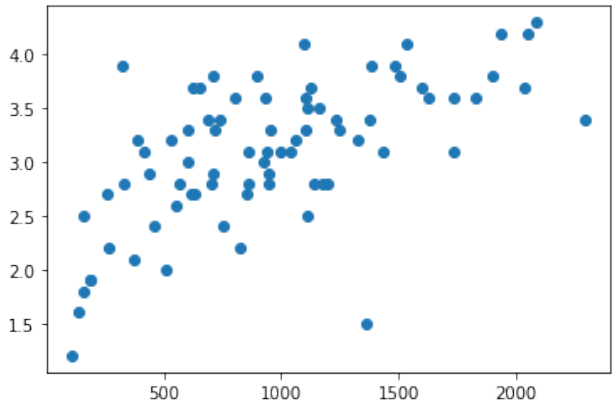

Fig. 1. The y axis indicates the weight of the games, the x axis indicates the number of sentences.

It is not the rule, but given the data, increasing the number of sentences usually increases the complexity of the game.

The luck feature indicates how much luck is present in the game, how much the game relies on luck and random events. This feature is very difficult to be computed precisely for a large quantity of generic rulebooks, thus this feature has been approximated as the number of sentences that contain luck features in the game's rulebook.
The main components that define randomness (luck) inside a board game are:

- shuffling cards;
- rolling dice/die;
- random events, usually game specific.

In this paper, the luck feature has been extracted in 2-phases. In the first phase, an AI model was trained on a specific dataset of around 120 sentences randomly extracted from all the rulebooks of the given first dataset (more on the datasets in the following chapter).
The sentences were tokenized and padded before giving them as input to the AI models.
The KNN, K-means, decision-tree approaches were not satisfactory, but some improvements were seen using decision-trees, for this reason the final model used for the first filtering of sentences is a random forest which is an ensemble of many decision trees, and it should gain more stability with the results.
However, the data from the random extracted sentences are not enough, the model performed quietly good on the samples, but very bad on the overall rulebooks, for this reason, a second phase has been introduced.
In the second phase another filtering technique is applied, this time to try to reduce the number of false positives, that is, sentences which are classified as "lucky" but they are not. The implementation of this second filter is based on keywords that indicate some sort of luck mechanic in the game.

These keywords are based on the three main randomness components that we can find in board games.

The last used feature is bookkeeping.
Bookkeeping indicates the technical skill needed in order to manage correctly resources in the game.
In a wider definition, bookkeeping can also take into consideration the managing of victory points, but it will complicate even more the analysis, so in this paper bookkeeping does not refer to victory points, but in-game resources.
A first approach was tried, using an embedder; the rulebooks were firstly embedded and then the similarity of the rulebooks with some bookkeeping articles was computed, but the results were very poor, the cause is given to the fact that an embedder works on the whole text, but the bookkeeping components are a few compared to the overall rules and elements of a game.
The approach also needed to be different from the one used with luck, because in the first case, luck can be easily extracted by analyzing single sentences, the same cannot be told for bookkeeping, which encompasses a wider argument.
For all these reasons a very simple approximation has been computed, keywords referring to bookkeeping have been extracted from articles, board games with a lot of bookkeeping mechanics and from a lot of field specific words.
These keywords, have been searched and counted in the rulebooks.
It might be an easy approximation, but many board games with a very present bookkeeping component like Terraforming Mars had a higher value with respect to other board games with little or no bookkeeping (exactly what I was looking for).

Extracting good features is hard, but how is it chosen that they are the best for the actual goal of the paper?
To answer this question, some regression models were trained and tested, the one with the most promising results was the gradient boosting model; This model was trained on all the possible combinations of features, and the ones that were able to achieve the lowest MSE were chosen as the best.

The figure 2. is used to simplify the understanding of the workflow used.

III. EXPERIMENTAL RESULTS

The analysis conducted in this paper refers mainly to the BGG database, however other two datasets that are linked to BGG have been used in order to have some specific sample to base the analysis.
The BGG dataset contains many features about the games as well as many important features and comments about the users of the games.
The whole paper is based on the BGG weight score, that has been extracted using BGG APIs, and some other features (previously discussed) have been extracted in the exactly same manner.
In order to have a sample of board games on which the rulebooks were extracted from other sources, external from BGG, another dataset has been used. This dataset called BoardGames.csv, contains the name, the ID and the name of the rulebook used for the analysis; these data were fundamental in order to link the BGG data with the rulebooks of the board games.
The games used as sample are 79, many from the top 100 scored games of BGG ranking.
The last dataset used is called random_sentences.csv and it is the one used for training and testing of the random forest classifier used in the luck extraction.
The random sentences are taken at random from the rulebooks of all the 79 boardgames, after the extraction of these sentences, some were discarded in order to keep the dataset balanced, the label for each element can be a 1 if the element contains some evidence of randomness or luck and 0 otherwise.

The metrics used for the training and testing of the regression models are the mean squared error (MSE) and the R-squared.

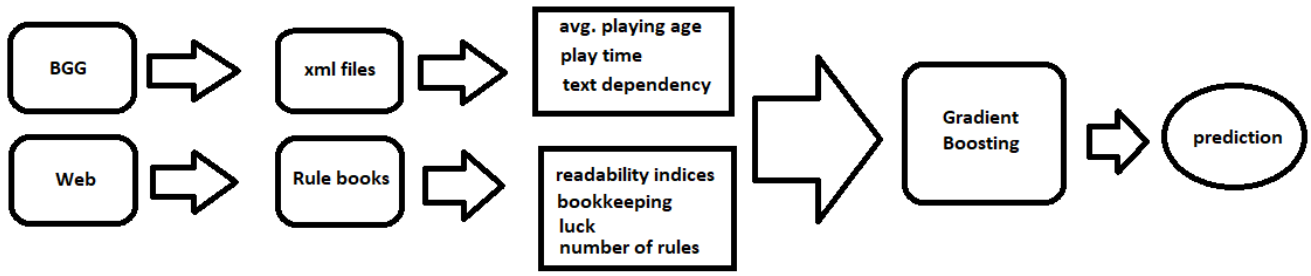

Fig. 2. Workflow currently explained

The mean squared error measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases.
The R-squared metric, tells us how much of the variation in the variable y can be explained by variable x.
The Gradient boosting model was able to achieve an MSE of 0.194623047691184 and an R-squared of 0.6601541914222643, considering the following features:

- Dale Chall readability;
- Gunning fog readability;
- luck;
- bookkeeping;
- average recommended playing age;
- dependency on text;
- playtime;

The following table represent some of other combinations of features with different final scores.

```
0: 'Forecast_readability'
1: 'Smog_formula'
2: 'Flesch_reading_ease'
3: 'PSK_readability'
4: 'Dale_Chall_readability'
5: 'Gunning_fog_readability'
6: 'number_sentences'
7: 'luck'
8: 'bookkeeping'
9: 'average recommended playing age'
10: 'dependency on text'
11: 'playtime'
```

Fig. 3. from numbers to features

Features used	MSE	R-squared
(2, 9)	0.3337	0.417
(2, 11)	0.3618	0.368
(3, 11)	0.2686	0.531
(0, 1, 8)	0.8384	-0.464
(0, 2, 9)	0.4003	0.301
(0, 3, 10)	1.4069	-1.457
(3, 4, 11)	0.2825	0.507
(4, 8, 9, 11)	0.2074	0.638
(3, 4, 8, 9, 11)	0.2139	0.627
(0, 4, 8, 9, 10, 11)	0.2290	0.600
(3, 4, 5, 8, 9, 11)	0.2259	0.606
(0, 4, 5, 7, 9, 10, 11)	0.2258	0.606
(4, 5, 7, 8, 9, 10, 11)	0.1946	0.660
(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	0.5921	-0.034

IV. CONCLUDING REMARKS

The dataset contained only a small amount of the overall board games available, the amount of possible choices that a player can consider during the gameplay was not considered, thus the prediction of the game complexity might be very wrong for games like chess and go as discussed previously.
Adding more board games and more data should lead to the overall best features used in the computation of board games complexity using the methods discussed in this paper.
Even if the dataset contained a small amount of data, the gradient boosting model was able to capture some dependencies between the features and obtain reasonable predictions.

For future work, using more data is advised, it is also possible to shift the analysis from the computation of complexity to the computation of learnability for the board games.
Learnability can be expressed as the time needed to learn the basic skills and mechanics of the game in order to play it. This feature can be used in real world applications in order to understand how much time a group of people or a single person will need to take into consideration before actually playing the game.
Learnability is also easier to compute, in fact, some features like bookkeeping, and amount of possible choices or luck, will not be taken into consideration since they are not required for the learnability purpose. Excluding the computation of the harder features, learnability should be a good subject for further and future studies.

REFERENCES

[1] A Data Driven Review of Board Game Design and Interactions of Their Mechanics — IEEE Journals Magazine — IEEE Xplore
[2] A Practical Approach to Language Complexity: A Wikipedia Case Study — PLOS ONE