Online Sorting Hat Quiz Analysis Using Machine Learning

Isabella Valdescruz
Northwestern University
Isabellavaldescruz2016@u.northwestern.edu

1. Overview

Harry Potter is a world famous name that has created a multi million dollar movie industry, theme parks, and tons of merchandise. The official fan website Pottermore.com, is a site that Harry Potter fanatics can create a log in and go through a simulation of what being a student in Hogwarts is like, a part of this simulation is the renowned Sorting Hat Quiz.

The four houses of Hogwarts, Gryffindor, Hufflepuff, Ravenclaw, and Slytherin are all extremely different and each one has its own characteristic traits that the Sorting Hat looks for in students in order to place them into the correct one. Fans of Harry Potter often relate to the values of one of the four houses while reading, so when this online official sorting hat quiz came out, many people were very excited. J.K. Rowling herself had a say in how the quiz sorted the people into the different houses.

For my project, I wanted to analyze this quiz and see if I could figure out what machine learning technique the algorithm implemented to deduce the correct house for a user. I then wanted to take it one step further and see if the models I created in Weka would hold when tested on actual Harry Potter character's themselves.

2. Data

For a training set, I found a data set on Pottermore.com, that had a total of 1,660 responses from users of the questions they answered as well as the house they were placed into. The quiz has several different versions; it has a total of 7 questions and 3-4 possible questions for each one. Regardless of version, however, each question had a certain theme in order to find out something deeper about a user's personal values and character traits.

I created an arff file from this data set, with each line simulating a user's responses from the quiz. I did find that there were duplicate responses in the data set, which makes sense since there were so many users and around 9,000 different versions of the quiz. At first, I took these duplicates out because I thought they would affect my outcome. After some thinking about it, however, I realized that the duplicate answers would actually help my model because it would give more data for it to draw conclusions from and thus I added them back in. I used ten-fold cross validation on this set to test it in Weka.

For my other testing set on characters, I simulated a quiz response set for five characters, Cedric Diggory, Harry Potter, Severus Snape, Lord Voldemort, and Hermoine Granger. I used data from the Harry Potter wiki that has a very detailed analysis of every main character in the books. I added the correct house at the end of the line and ran each character individually on the training set to see if it outputted the correct classification.

3. Results

I decided to use J48 (Decision Trees), IBk (K-Nearest-Neighbor), Naïve Bayes, and MultiLayer Perceptron as possible classifiers for the data. I used Decision Trees because I was interested in seeing which variable gave the highest information gain in the data set and I thought that IBk and the MultiLayer Perceptron would make sense because they include weights on variables. I hypothesized that the online quiz must take weights into account because there are some traits that are unanimous for certain houses, occur in other houses, and do not exist in the others.

My results from cross-validating on the training set was that K-Nearest-Neighbor performed best out of the four classifiers I tried, closely followed by the MultiLayer Perceptron.

Classifier	Accuracy
J48	92.801%
KNN	95.4023%
Multilayer Perceptron	95.0988%
Naïve Bayes	72.9583%

My results from my testing set that I created from the characters matched up with this data as well. For all five characters, K-Nearest-Neighbor and the MultiLayer Perceptron gave a 100% accuracy rate, correctly classifying each character into his/her correct house. Naïve Bayes and J48 sometimes got it correct, but placed some characters into a house that they were close to but not a perfect match with.

4. Conclusions

These results make me conclude that the sorting hat quiz uses an algorithm that implements weights paired to each of the houses. For future testing I would like to combine the results from the J48 model that listed the attributes that gave the highest information gain and then weight the attributes accordingly and perform a linear regression. This is how I think I would be able to achieve perfect results as the model wouldn't get confused with things such as overfit. See the individual character results below.

Harry Potter - Gryffindor

Classifier	Classified House
J48	Slytherin
Naïve Bayes	Gryffindor
KNN	Gryffindor
Multilayer Perceptron	Gryffindor

Hermoine Granger - Gryffindor

Classifier	Classified House
J48	Gryffindor
Naïve Bayes	Ravenclaw
KNN	Gryffindor
Multilayer Perceptron	Gryffindor

Lord Voldemort - Slytherin

Classifier	Classified House
J48	Slytherin
Naïve Bayes	Slytherin
Multilayer Perceptron	Slytherin

Severus Snape - Slytherin

Classifier	Classified House
J48	Gryffindor
Naïve Bayes	Slytherin
KNN	Slytherin
Multilayer Perceptron	Slytherin

Draco Malfoy - Slytherin

Classifier	Classified House
J48	Slytherin
Naïve Bayes	Slytherin
KNN	Slytherin
Multilayer Perceptron	Slytherin

Cedric Diggory – Hufflepuff

Classifier	Classified House
J48	Hufflepuff
Naïve Bayes	Gryffindor
KNN	Hufflepuff
Multilayer Perceptron	Hufflepuff