Mandatory Individual Assignment - Data Mining

by Lucas Klafke Beck

April 12, 2017

Questions

The questions explored by this report in the data set are presented below:

- Is it possible to determine the student degree based on his/her other interests?
- Can we cluster students in a meaningful way in terms of their gender by looking into their available physical characteristics (shoe size and height)?
- Can we find patterns on the games the student have played (e.g if played Fifa, also played counter strike)?

Data preprocessing

Whenever loading the data, some precautions were taken for all the attributes. For non-numeric attributes which had open text entries, numeric characters were removed, case was disregarded and all the possible valid answers were considered. For instance, in gender all "m", "man" and "male" were considered to be valid.

For physical characteristics numeric fields, some constraints were also established. Non-numeric values were removed and physical constraints added. For instance, height was only considered if between 55 and 280, based on historical data.

Filling missing values

After the initial parsing, some values were missing for height and shoe size. Thus, the median value within gender for those attributes was used for all tuples who had missing values. For instance, if a tuple had gender male and missing height, the median of the height across all male students was used. This was done given that this attributes are used to cluster according to gender (intended result).

This bias the data towards the intended result. However, with missing data some compromises have to be taken one way or the other. Moreover, since the

number of data points with missing values was small, the effect of the bias was also.

Ignored tuples

Students that marked "none" for games were removed from the Apriori analysis since they were not relevant for this analysis. Students who had no valid gender were removed from the clustering analysis.

Normalization

All numeric values were normalized in values between 0 and 1 using the min-max normalization technique.

Analysis and Results

Student degree based on interests

For the K-nearest neighbors, the euclidean distance was used in order to determine similarity between tuples. Since all the attributes concerning students interest had an order starting in "not interested" and ending in "very interesting", each possible value was given a numeric equivalent between -0.5 and +0.5. The values were equally distributed in this scale. That way, a student having "not interested" is closer to a student having "sounds interesting" than to another having "very interesting". Moreover, 2/3 of the data were used as training set and the rest as the test set. The data was also randomized before every run.

In order to evaluate how good the algorithm is in answering the question, the measure used was accuracy. The reason is that there is more than one category that a student degree may fall in. Thus, it is not possible to use sensitivity or specificity given there are no positives or negatives in this case. After running the algorithm 200 times and taking the average for neighbors from 1 to 50, the following accuracy plot was generated As shown in figure 1. The best k for the model is at $k = 7$ and has 37% accuracy. This shows that using only the interest of students does as independent variables, does not give us a good prediction tool for the degree. Even further, The degree which the student is in, does not tell us which topics he/she is interest within data mining.

Clustering students on height and shoeSize

The clusters are represented in figure 2. As expected, the clusters do not fully represent the gender. The algorithm was trained using only age and shoe size data, and even though it is known that man in general have higher measures than women, there are exceptions. However, from the 11 girls on the dataset only one (9%) was classified in a different cluster. For man, around 18 % stayed out of the male dominant cluster. Therefore, it is possible to conclude that

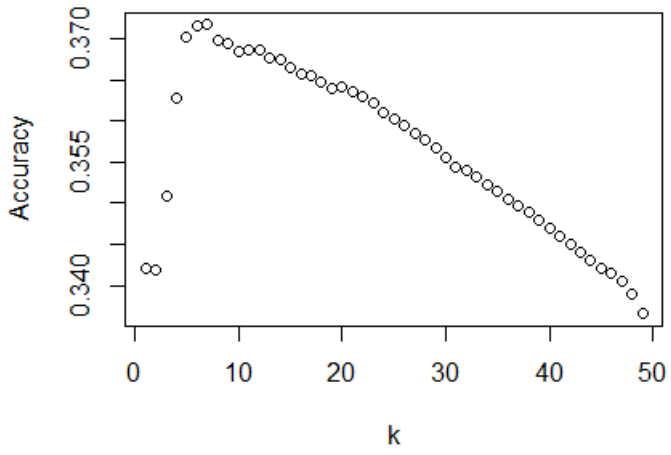

Figure 1: k vs Accuracy

clustering students by shoe size and height with k = 2 gives an good approximate distinction across gender among clusters.

Figure 2: Clusters representation

Pattern Mining on Games

When looking at the patterns within the game sequence, I was interested in those rules with high confidence. Thus, I could see if playing certain game(s) would most likely mean that you play some other(s). Thus, I set the confidence with a high threshold of 85 % and a relatively low support of 14 % . This means that catching games that may not occur much, but when they do, in 85% of times they will lead to other games. My finds are presented below:

battlefield 4 $\rightarrow$ counter strike go
counter strike go, wordfeud $\rightarrow$ minecraft
fifa 2017, candy crush $\rightarrow$ angry birds
stanley parable $\rightarrow$ minecraft
farm Ville, candy crush $\rightarrow$ angry birds