Data Mining – Individual

Christoffer Horn - horn@itu.dk

Total unit count: 5249

Introduction

During the first lecture all students in the course answered a questionnaire. This study is based on data obtained from that questionnaire. The questionnaire was answered by 75 students.

This study has looked into three research questions. The three research questions are

Classification using supervised learning
Is it possible to classify the students' degree based on their interests, preferred phone OS and games played with a correctness of at least 60%?

Frequent pattern mining
Which programming language combinations are the most common on this course?

Clustering
What is the gender of a student given a height, age and shoe size?

Preprocessing

The preprocessing methods applied were data cleaning and normalization.

Data Cleaning

During the preprocessing no data points were discarded. Instead it was chosen to fill in missing values or unrealistic outliers with a default value.

Normalization

The numeric data that needed normalization includes height, age and shoe size. No other numeric data was used in this study. The data was normalized to the interval going from 0 to 1. This was done to reduce the height's influence in the Euclidean distance.

Classification

For classification of the students the k-nearest-neighbors algorithm was used. The attributes used for classification are

1. how interested they are in database design,
2. how interested they are in predictive models,
3. how interested they are in grouping similar objects,
4. how interested they are in visualization,
5. how interested they are in finding patterns in sets,

6. how interested they are in finding patterns in sequences,
7. how interested they are in finding patterns in graphs,
8. how interested they are in finding patterns in text,
9. how interested they are in finding patterns in images,
10. how interested they are in coding data mining algorithms,
11. how interested they are in using off the shelf data mining tools,
12. phone OS used, and
13. the games they have played

The attributes 1 through 11 are given a value from 0 to 3 based on their answer. The answer "Not interested" results in a 0, "Meh" results in a 1, "Sounds interesting" results in a 2 and "Very interested" results in a 3.

The distance between two data points, p1 and p2, is calculated using the following formula:

$$dist(p_1, p_2) = \left|p_{1_{games}}\ not\ in\ p_{2_{games}}\right| + \left|p_{2_{games}}\ not\ in\ p_{1_{games}}\right| + \left|p_{1_{phone\ OS}}\ not\ in\ p_{2_{pho\quad OS}}\right| + \sum_{i=1}^{11} p_{1_i} - p_{2_i}$$

The results from the classification for different k values can be seen in Figure 1. Additionally the classification for $k = 6$ can be seen in Table 1. Since most of the students are design track students the algorithm will classify most students as design track students for high k.

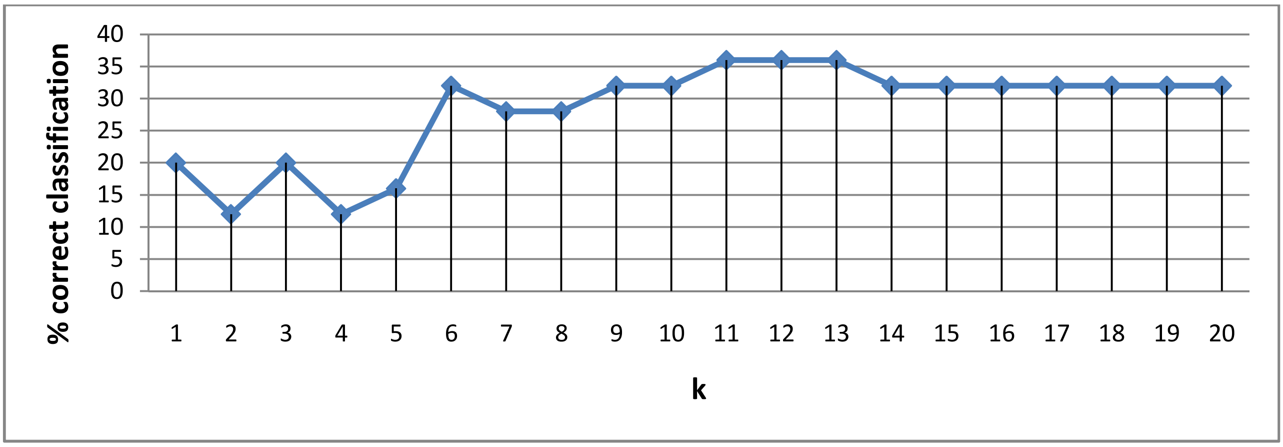

Figure 1: The percentage of correct classification for differen k.

Frequent Pattern Mining

The frequent patterns found are shown in Table 2. The association rules found are shown in Table 3. The implementation only supports finding association rules for one programming language implying another. It does not support multiple programming languages implying multiple programming languages. It should be noted that the association rules with Java in them might not be true, since Java is known by almost all students. So association rules without Java are more interesting e.g. those students who know C# also know C++ and vice versa.

Correct classification	Assigned classification
SE	DT
SE	DT
SE	DT
AC	games
DT	DT
DT	DT
SE	DT
games	DT
SE	DT
SE	DT
games	AC
DT	DT
DT	DT
DT	DT
SE	DT
SE	games
guest	DT
SE	DT
DT	DT
DT	DT
DT	games
SE	DT
SE	DT
SE	DT
DT	DT

Table 1: The classification of students using $k = 6$.

Pattern	Support
Java	93%
C#	43%
Python	32%
C++	31%
Java, C#	41%
Java, Python	29%
Java C++	29%
C#, C++	28%

Table 2: The frequent patterns found using Apriori and a support threshold of 30%.

Association Rule	Support	Confidence	Correlation
Java => Python	29%	31%	0,62
Java => C#	41%	44%	0,71
Java => C++	29%	31%	0,71
Python => Java	29%	92%	0,62
C# => Java	41%	97%	0,71
C# => C++	28%	66%	0,78
C++ => Java	29%	96%	0,64
C++ => C#	28%	91%	0,78

Table 3: The association rules found by Apriori with a support threshold of 30%.

Clustering

The clustering was done using the K-means algorithm.

It was not possible for me to get the clustering to work within the given time limit. The clustering does not terminate, even though it looks to have settled on a clustering. The clustering was run with $k = 3$ since I have decided to have three genders in the data set: male, female and apache helicopter as a default value. The algorithm could not find a good clustering where all males, females and apache helicopters got assigned to different clusters mostly containing similar gender students. Using some other distance measure might make this possible.

Conclusion

From this it can be concluded that it is not possible to classify students' degree based on their interests, phone OS and games played, at least not with the implemented distance measure. Another distance measure might make it possible.

Additionally, it was found that Java is by far the most common language known by students on this course. A popular combination of languages is C# and C++, since 91% of people knowing C++ also knew C#.

Finally, it was found that it was not possible to cluster students into genders based on age, height and shoe size using the implemented distance measure.