Data Mining:
Individual project

Anders Wind Steffensen (awis@itu.dk)

Spring 2017
17th of April

This report contains a total of 4792 characters, including titles, figure text and text in tables, excluding this front page.

Introduction

This report and implementation focuses on the implementations of the algorithms. They should be generic and able to handle any data points given to them if they conform to certain standards. Furthermore, for classification and clustering two different implementations have been given to be able to assess and compare their effectivity and correctness.

The implemented algorithms and the hypothesizes are the following:

- K-Nearest Neighbors and ID3
  - *Can the gender of a participant be predicted from a data point?*
- Apriori
  - *Is it possible to recognize patterns in which programming languages the participants are proficient in?*
- K-Means and K-Medoids
  - *Does data points cluster on which degree the participant is pursuing?*

The program can be run from the main of `ProjectRunner` class and every result shown in this report has been created using seed value 15 which can be provided as the first command line argument.

Pre-processing

Several techniques, such as outlier removal and normalization are used in the preprocessing phase.

Certain rules have been put up for the attributes of the data points. For example, for the numeric attribute “age” the value must be between 18 and 65 to be able to handle faulty entries (one entry had Age=999 which does not seem likely). If the data point does not abide the rules, then the data point is not used. This is a very strict policy and could entail skewed results.

Furthermore, most multiple choice answers have been converted to Enums and the description given in the “other” option have been discarded. For the given dataset, the approach seem fine since only two entries had entered values in the “other” option.

The “Proficient Language” attribute had no premade choices so to match the same answer written different ways; multiple strings map to the same result. The “Gender” attribute has been matched to a binary nominal attribute for either Male or Female which matches all the answers.

The three data types which consists of multiple answers, proficient languages, commute means, and games played are also represented in another aggregated attribute describing the amount of these values.

Both the originally only numeric value “Age” and the aggregated attributes are normalized after the outliers have been removed to ensure less skewed results.

Most of the data cleaning can be seen in the class `AnswerDataLoader` or `Normalizer`.

Classification

To be able to classify a data point a specific attribute is picked as the classification attribute. To compare and assess KNN and ID3 the gender attribute has been picked. Since it is a binary nominal value a confusion matrix can be created for both methods. All attributes are used for both methods and a training sample size of 45, and test size of 20 is used.

	K-Nearest neighbors	ID3
True positives(female)	1	0
False positives	2	3
True negatives (male)	17	17
False negatives	0	0
Accuracy	90.0%	85.0%
Error rate	10.0%	15.0%
Sensitivity	33.3%	0.0%
Specificity	100%	1.0%
Precision	33.3%	0.0%
F-score	33.3%	NAN

It does not seem that ID3 can categories data points as females, but by picking other sample sizes and different seeds it did show some flexibility and ability. Generally, K-nearest neighbors performs the best but it is difficult to evaluate with the small sample sizes.

Pattern mining:

Apriori is implemented to do pattern mining. Specifically, a support threshold of 8 and a confidence threshold of 84% is used. All data points and all attributes of the data points are used.

- F# -> Java, C#: confidence: 84%
- C, F# -> Java, C#: confidence: 88%
- JavaScript, C++ -> Java, C#: confidence: 88%

These values have a high support in the data set and furthermore a high confidence which does seem to entail some set patterns in proficient languages.

Clustering

All data is used but for K-Means all sequence-attribute are omitted due to difficulties in calculating “mean” values for sequences. To measure, compare and answer the hypothesis, the percentage of the most common “degree” of a cluster is calculated. The number of clusters was decided by experimenting and k=4 seemed to provide the best results.

Cluster (most common degree)	K-Means	K-Medoids
SDT-DT	65.2%	38.8%
SDT_DT	57.1%	57.1%
SDT_SE	70%	44.4%
Games-T	66.6%	50.0%

The two methods seem equally bad at clustering the data points based on this attribute, but this is probably due to the high dimensionality of the data points and might do better if only the “Games played”, “Proficient Languages” and “Age” attributes were used.

Conclusion

Generally, the algorithms do seem to pick up some patterns, but they struggle due to the low volume of data and its high dimensionality. One could reduce the number of dimension by looking at correlations among certain attributes and then use only those. To accumulate more data, several years of questionnaires could be combined.