Advanced Topics in Machine Learning Programming Task

Jannis Becke, Christian Lausberger and Maik Riestock

Otto von Guericke University Magdeburg - Data and Knowledge Engineering Group

1. APPROACH

an der aufgabenstellung angelehnt

- analyse the data set and its various attributes
- clean the data (e.g. missing values)
- select an appropriate subset of the attributes and explain your choice
- use different suitable machine learning algorithms (either implement them, or use existing libraries, e.g. Weka)
- determine the quality of your model (e.g. through cross-validation, log loss3, confusion matrix)
- compare your results between different algorithms

  fr die bearbeitung dieser aufgaben verwendeten wir die libary von weka.
  weka ist eine

2. DATABASE

source what it is about what the attributes are about

3. PREPROCESSING

3.1 attributes

remove of attributes

3.2 instances

handling of missing values
  remove of instances?

3.3 conclusion

4. CLASSIFICATION

Was istr das problem? 3-class problem
  welche vorgaben? Specifically, different suitable machine learning algorithms should include:

- algorithm based on SVM
- algorithm for semi-supervised classification
- one additionally supervised algorithm

  overview of: used classifier used evaluation methots
  results contains:

- used data set
- used classifier(with configuration)
- used evaluation method(with configuration)
- numbers(correct in generell, for each class)

Table I. Ergebnisse der Klassifikation des IBK und RandomForest

Set	Algorithm	Evaluation	Result
set	SVM%	Percentage Split%	52%

explain the results

5. CONCLUSION

List of Tables

APPENDIX

A.   TASK DESCRIPTION

Use the Diabetes 130-US hospitals for years 1999-2008 Data Set. This dataset contains records with 55 features for more than 100000 patients. 55 features include information about the diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information. Note that even though some attributes are codified using numeric values, they are nominal (not numeric) attributes. Browse additional information about the attributes in the paper cited on the dataset website. The goal of this assignment is to find suitable methods in the area of machine learning to determine the readmission attribute of a patient and to estimate the quality of selected approaches.

In order to achieve this goal, the following tasks have to be completed:
- analyse the data set and its various attributes
- clean the data (e.g. missing values)
- select an appropriate subset of the attributes and explain your choice
- use different suitable machine learning algorithms (either implement them, or use existing libraries, e.g. Weka)
- determine the quality of your model (e.g. through cross-validation, log loss3, confusion matrix)
- compare your results between different algorithms

Specifically, different suitable machine learning algorithms should include:
- at least two classification algorithms, one of which, SVM, you learn during the course
- at least one algorithm for semi-supervised classification. Use a part of the dataset as unlabeled data for learning (omit the label)