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 folgende libs:
weka - provides a lot of basic functionality like data storage,filtering, visualisation and additionally many classifier and clustering algorithms
collective classification - provides semi-supervised algorithms
libSVM - provides additional SVM algorithms

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? minimum

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

overview of: used classifier

- NaiveBayes
- DecisionTable
- LibSVM
- YATSI

zur eveluirung der trainierten classifizieren verwendeten wir die beiden blichen evaluirungs methoden: **Percentage Split** - used with 80% as training set and 20% as test set
Cross-Validation - used with 10 folds

whrend der durchsicht der ergebniise ist uns aufgefallen dass die ergebisse bei vielen classificatoren alle instanzen in eine classe classifizierten. das problem entsteht daraus dass die anzahl der insanzen zwischen der classen stark variiert. UM dieses problem zu reduzierten haben wir zwei anstzen ausprobiert die in section One Class Problem vorgestellt werden.

Table I. Classifier result for each evaluation method

Algorithm	Evaluation	Correctly Classified
NaiveBayes%	Split%	57.32%
NaiveBayes%	Cross-Validation%	57.14%
DecisionTable%	Split%	57.53%
DecisionTable%	Cross-Validation%	57.75%
SVM%	Split%	-%
SVM%	Cross-Validation%	-2%
YATSI%	Split%	-%
YATSI%	Cross-Validation%	-%

results mit alle in 1 klasse
explain the results

5. ONE CLASS PROBLEM

vieler klassifikatoren classifizierten alle instanzeni in eine klasse. dieses problem entsteht wenn die anzahl der instanzen der klassen sich stark unterscheided. um dieses problem zu lsen haben wir zwei ansttze verwendet die im folgen erlutert werden

5.1 Even Classes

ansatz durch zustzliches preprocessin
bei diesem ansatz haben wir die anzahl der instanzen der jeweiligen klassen angeglichen. dafr haben wir als erstes die klasse mit der genigsten anzahl von instanzen ermittelt. danach haben wir bei den anderen klassen durch zufllige auswahl soviel instanzen gelscht bis die klassen angeglichen waren.

mit diesem neu erstellten datenset haben wir nun unsere classifikationen erneut gemacht. jedoch steltle sich herraus dass nun die klassifizieren die klasse mit den ehemaligst wenigesten instanzen bevorzuge und nun alle instanzen darin klassifizierte.

This documentation was created in the context of the course Advanced Topics in Machine Learning summer term 2014/15. This course was held by: Prof. Dr. Andreas Nurnberger, M.Sc. Tatiana Gossen; Research group Data and Knowledge Engineering Group, Otto-von-Guericke-University of Magdeburg, Germany.

5.2 2 Class Problem

ansatz durch vernderung des classification attribut und damit der classification//

6. 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)