CIND119 Midterm paper
Project Description - The Ask

Author: Mori Fihri

Project Goal:

Build a classification model based on the training data to predict if a new customer is approved or not.

Project Description

Your task in this project is to work with a team of data scientists and solve a problem based on a given dataset for an organization. There are multiple datasets available for this project. These datasets are described at the end. You will have to choose only one of the dataset for analysis for your project. Your team will consist of 4 to 5 students. Each team will submit one project report and
will give a 10 minutes presentation (using PowerPoint slides).

In your project, you will perform the following steps on the selected dataset:

1. Data Preparation
2. Predictive Modeling/Classification
   a. Classification using Decision Tree
   b. Classification using Naive Bayes
   c. Choose one classification algorithm of your own, apply and describe it
   d. Compare the results of the 3 techniques
3. Post-prediction Analysis
   a. Apply cluster analysis / association rule mining on the results of the classification to provide customized recommendations to the organization for its customers.
4. Conclusions and Recommendations

Describe your results in your project report for each of the above steps. These steps are further explained in the sections below. The suggested tool for this project is Weka but you can use any other tools as well.

Data Preparation

The first and foremost step of data mining process is to understand the data and identify the research question(s). Here are some suggestions to explore and understand datasets:

Look at the attribute type; e.g., nominal, ordinal or quantitative.

Find max, min, mean and standard deviation of attributes.

Determine any outlier values (records) for each of the attributes or attributes under consideration (min, max, std. dev, scatter plots, box plots or others can be used).

Analyze the distribution of numeric attributes (normal or other). Plot histograms for attributes of concern and analyze whether they have any influence on the class attribute.

Load the dataset in Weka and click on visualization tab. Which attributes seem to be correlated?

Which attributes seem to be most linked to the class attribute?
Which attributes do you think can be eliminated or included in the analysis?
Determine whether the dataset has an imbalanced class distribution (same proportion of records of different types or not).

Determine whether you need to handle missing values or transform any attributes (e.g., by normalizing the attributes, discretizing numeric attributes to categorical attributes, etc.). Weka filters (on the main tab) can be used for this purpose. Describe your findings for data preparation in your report.

Predictive Modeling (Classification)

Now apply the classification algorithms, Decision Tree and Naïve Bayes, which you studied in the course on your dataset. Also, choose a classification algorithm of your own choice, explain it a at a high level and compare your results.

- You will predict the class attribute by using each classification algorithm.
- Determine the right strategy for dataset split: simple training or testing, 10-fold cross validation, 3-fold cross validation, etc.
- Investigate the use of different parameters present in Weka for Decision Tree and compare your results obtained in different settings. Understand your decision trees generated by Weka.
- Repeat the same process for Naïve Bayes and the third classification algorithm of your choice.
- Determine your performance measures (accuracy, recall, etc.).
- Identify which algorithm performs well and in which settings.
- Describe results of predictive modeling in your report. Explain your interpretation of output of each of the algorithm in your report (e.g., explain decision tree). Explain the criteria you used to select the best performing algorithm (e.g., accuracy, true positive, recall, comprehensibility, etc.).

Post-predictive Analysis

You have identified the model to predict the type of customer based on your selected dataset. Now you need to identify the characteristics of these customers that you predicted as important (e.g., churning customer, prospective subscribers, prospective borrowers, etc.) based on the problem for your dataset. You can apply clustering analysis or pattern mining (association rules) or both to find the characteristics of these customers. In a real world scenario, you may need to perform additional analysis on the output of interest of the prediction algorithm to identify and understand different groups/segments and link them to the business objectives.

To make it easy for yourself in this project, you can filter out the records of the class from the original dataset (not based on the predictions of your models) that you consider not interesting for the business (e.g., not churn) and keep only the class that you would like to analyse further. For example, if you want to analyse the people that are likely to churn, just take all the record from the original dataset that have TRUE as a true class label (that is ignore the prediction of the model).

Using the filtered dataset, you can now apply clustering or pattern mining to conduct further analysis and suggest your recommendation.

Here are some suggestions for clustering:

- Investigate the use of K-Means algorithm to segment the data of the predicted class of importance.
- Analyze each segment (group or cluster) and identify the characteristics of customers (type of records) in each group; e.g., the characteristics of a group/cluster can be determined by finding the majority of attributes in that group.
- Explain your interpretation of characteristics and state the recommendations for the organization.

Here are some suggestions for pattern mining:

Explore association rules based patterns for the records of the class of interest by using the Apriori algorithm in Weka on your dataset.

- You may have to use selected qualitative (*categorical or ordinal)* attributes to discover patterns.
- Try different values for minimum support and confidence, select the values that provide the appropriate number of rules and justify your selection.

- Identify the frequent and logically correct patterns and state your recommendations for the organization on different types of customers belonging to the predicted class.

- Describe your analysis for this section in your report.

Conclusion and Recommendations

State your major findings from different sections. State your recommendation to the company that they can put into place to solve their problem.

Credit Card Dataset

In order to provide loans to customers, a bank needs to make right decision in determining who should get the approval and who should not. This dataset is the German Credit Data that contains 20 attributes and the class attribute showing a good or a bad credit risk. Your team of data scientists will need to develop a data analytics based strategy for the bank managers that can help them in making a decision about loan approval for the prospective applicants.

1. Creditability: The class attribute (qualitative)showing whether the credit rating is good or bad.
2. Account Balance: Checking account status (1: < 0 DM, 2: 0 <=...<200 DM, 3 > 200 DM, 4: No checking account), where DM= Deutsche Mark (qualitative attribute).
3. Duration of Credit (month): Duration of credit in months (numerical)
4. Payment Status of Previous Credit: Credit history (qualitative) 0: no credits taken, 1: all
credits at this bank paid back duly, 2: existing credits paid back duly till now, 3: delay in paying off in the past, 4: critical account.
5. Purpose: Qualitative attribute showing the purpose of the loan (0: New car, 1: Used car , 2:
Furniture/Equipment, 3: Radio/Television, 4: Domestic Appliances , 5: Repairs ,6: Education ,7:
Vacation, 8: Retraining ,9: Business, 10: Others)
6. Credit Amount: Numerical value showing the credit amount
7. Value Savings/Stocks: Qualitative attribute showing average balance in savings and stocks (1 : <
100 DM, 2: 100<= ... < 500 DM, 3 : 500<= ... < 1000 DM, 4 : =>1000 DM, 5: unknown/ no savings
account)
8. Length of current employment: Qualitative attribute showing length of employment (1 :
unemployed, 2: < 1 year, 3: 1<=...<4 years, 4: 4<=...<7 years, 5:>=7years).
9. Instalment percent: Installment rate in percentage of disposable income (numerical)

10. Sex & Marital Status: Qualitative attribute showing gender and marital status (1: male :
divorced/separated, 2: female : divorced/separated/married, 3 : male: single, 4: male :
married/widowed, 5 : female : single)
11. Guarantors: (Qualitative) Guarantors and co-applicants: (1 : none, 2 : co-applicant, 3 :
guarantor)
12. Duration in Current address: Qualitative value showing the duration in current address (1: <= 1
year, 1<...<=2 years, 2<...<=3 years, 3:>4years)
13. Most valuable available asset: Qualitative attribute showing valuable assets (1 : real estate
2 : savings agreement/ life insurance, 3 : car or other, 4 : unknown / no property)
14. Age (years): Numerical value showing age in years.
15. Concurrent Credits: Installment plans (1 : bank, 2 : stores, 3 : none)
16. Type of apartment: Type of housing (1 : rent, 2 : own, 3 : for free)
17. No of Credits at this Bank: Numerical value showing number of existing credits at the bank
18. Occupation: Job (Qualitative) (1 : unemployed/ unskilled - non-resident, 2 : unskilled - resident,
3 : skilled employee / official, 4 : management/ self-employed/highly qualified employee/ officer)
19. No of dependents: Numerical value showing number of dependents
20. Telephone: Qualitative attribute for telephone number (1: yes, 2: No)
21. Foreign Worker: Qualitative attribute showing whether the person is the foreign worker or not (1:
yes , 2: no)

How to Compare Your Classification Models

In order to evaluate and compare machine learning models with different features, a known approach is to create a baseline model first. This can be done by training one model (e.g., decision tree) on the training set using the entire feature set (all attributes) and evaluating its performance using the selected metric (such as accuracy, true positive rate, false positive rate, etc.) on the validation set (or test set).

Optimization of the model parameters or selection of features could be done by changing one variable (e.g., one parameter or one feature) at a time
and re-training the model on the same training set. Finally, one needs to compare the performance of different models built using different features on the same validation set (or test set). This would give an indication whether that variable (feature or parameter) has increased or decreased the performance.
In particular, here are two options for data splitting and evaluation. One needs more work in splitting and another one is straight-forward using Weka:

1. Using 3-ways data splitting (Here, you will have to divide a dataset yourself):

• Training (e.g., 60%): for training the model
• Validation (e.g., 20%): for evaluating and comparing the performance after varying
parameters, features, etc. **In addition, this is used for selecting the "best" model.**
• Testing (e.g., 20%): only used at the end to evaluate the final performance and report
the results of the selected models (best performing models from the above step)
• This video describes this approach:
https://www.youtube.com/watch?v=4wGquWGvGw

2. Using 10-fold cross validation (10-FCV) on the entire dataset:

• Create the base model by train and evaluate its (average) performance using 10-FCV.
• Change model parameters or features, retrain and re-evaluate the (average) model
performance using the 10-FCV strategy.
• Report the results of best performing models (using the best parameters and features
select from the above step).
Note: When you are splitting manually, always make sure stratified sampling is being used.
Weka applies stratified sampling by default in splitting datasets (e.g., using cross validation).

Project's Evaluation Criteria

Each group will be evaluated based on its analysis of the project as well as by the peers in the classroom. All the individuals will be evaluated based on workload distributions among the groups.

Please use the project template to submit one group report including workload distribution