Data Mining

Lab Exercises, WS 2016/17

Predictive modeling

Purpose

The purpose of this study is to learn how data mining methods / tool (IBM SPSS Modeler) can be used to solve predictive modeling tasks. The task consists in building classification models based on real data from industry.

The study will illustrate stages of data mining process, as guided by the CRISP-DM methodology:

- Initial exploration of data in order to understand data / discover inconsistencies,
- Preparation of data for modeling,
- Fitting a predictive model to data,
- Assessment of predictive performance of the model.

Input data

Data set: `copper_wire`

This data set describes quality of a batch of copper wire produced by a copper production plant. The variable `quality` represents quality of a section (roll) of copper wire produced. The remaining variables represent data from the production process and were recorded by the production monitoring system (these variables include temperatures, levels of various impurities in copper etc.).

The values of `quality` ≤ 6 denote good quality and the values > 6 denote poor quality of wire.

Tasks

The task involves:

- Building a model for automatic prediction of quality (good / poor) of sections of copper wire based on parameters of the production process.
- Fine-tuning the model (trying to maximize the overall profit defined in the Deliverables section below). The methods to be tried for model fine-tuning are described below.
- Assessing predictive performance of the model for new data.

The task will be solved using IBM SPSS Modeler (Clementine) data mining tool. A stream built in Modeler should realize the following stages of the data mining CRISP process:

- Connection to the input data / data sampling,
- Explanatory analysis of the data,
- Preparation of data for modelling,
- Building the predictive model,
- Assessment of predictive performance of the model.

Connection to the input data

The Sources palette contains nodes used to connect to source data in various formats.

Nodes to use in this task: Variable file node and/or SAS import node.

Exploration of the data

Useful nodes to visualize / summarize data:

- Data Audit node
- Plot node
- Distribution node
- Histogram node
- Table node
- Quality node

Preparation of data for modelling

Operations at the record level (Record palette)

- Select node – to subset data
- Sample node

Operations at the column level (Field palette)

- Type – to assign proper measurement level and role of columns in the process of modelling
- Filler node – to impute missing values, remove/coerce outliers, etc. (Note, if you want to impute column mean in place of a missing value (see Filler node), then you will also need to use Set Globals node, to pre-compute the column mean).

Building a predictive model

Sample nodes to try for classification (Modelling palette):

- Neural net
- C5.0 and/or other tree nodes
- Logistic regression
- SVM
- k-nearest neighbours
- …

Assessment of predictive performance

Nodes useful for obtaining standard and custom measures of model performance (Output palette)

- Analysis node (use User Defined Measure to compute the overall profit)
- Matrix node

Fine-tuning the predictive model

In order to maximize the overall profit related to classification of copper wire (or in other words – to minimize cost of misclassification events), you can try the following modifications of the stream:

- Fine-tune missing value imputation / outlier removal procedures – use supernodes created by the Data Audit node, or Filler node.
- Use different predictive modelling nodes. You may try the following nodes: decision trees, SVM, neural networks, logistic regression, nearest neighbours method, discriminant analysis.

- Experiment with specific settings of classifiers, esp. settings related to complexity vs. simplicity of the models (such as the regularization parameter C in the SVM model; leaf size and pruning parameter in tree algorithms; number of neurons in the hidden layer). You may also try model specific setting such as e.g., misclassification cost matrix in tree training nodes or bagging or boosting models ensembles (available in neural net node).
- Use feature selection. Following nodes can be used for feature selection:
    - PCA node (attempt to classify your data based on the set of first principal components rather than based on all original inputs),
    - Feature selection node (remove inputs not related to the target variable),
    - Alternatively, you can use feature selection capabilities offered by some predictive models (such as e.g., tree or logistic regression).
- Try to balance your data in terms of the proportion of class 0 versus class 1 in the training data (using the Balance node).
- Fine-tune sensitivity versus specificity of the model (by using the confidence column generated by the scoring node as the basis for the final classification decision).

Deliverables

- Scored data set (subset of copper_wire data set not used for model training),
- Measures of performance of the classifier:
    - Coincidence matrix,
    - Overall profit (measure aggregating the profit / cost related to good→poor and poor→good misclassification).

      Assume the following profit matrix for the classifier's decisions:

	Predicted as good	Predicted as poor
Good	100	50
Poor	-1000	50

Try to maximize the overall profit measured for the test partition.
Produce a report summarizing performance of your classifiers observed for different experiments related to model fine-tuning.
Attach the Modeler stream (.str file) to your report.