Institut für Informatik und Wirtschaftsinformatik -- Lehrstuhl für Wirtschaftsinformatik V
Dr. Heiko Paulheim / Oliver Lehmberg / André Melo

Exercise 5: Data Mining II – Time Series

Analyzing the Shopping Sequences Data Set

1. The Shopping Sequences data set is provided as on the website. Load the data set into RapidMiner. In the import dialog, set the data type of the CustomerNo and the Sequence attribute to integer. All other attributes must have the data type binominal (otherwise the operators cannot handle the data.

2. Mine sequential patterns from the data set using the GSP operator (min support = 0.3, max-gap = 1.1, window size = 0, min-gap = 0, positive value = 1). What can you conclude from the patters about the shopping sequence of people who buy an Asus EeePC netbook? Do you think that the initial performance of the netbooks fulfills the customer’s expectations?

3. Do your conclusions become even cleaner when you set window size = 2? Why?

4. Set min support = 0.2. What can you say about the shopping sequences of people who buy a HP Laserjet printer?

Analyzing the Travelers Data Set

1. The Travellers data set is provided as an Excel on the website. Load the data set into RapidMiner. In the import dialog, set the data type of the Person and the Sequence attribute to integer. All other attributes must have the type binominal (similar as above).

2. Mine sequential patterns from the data set. What can you conclude about the popular routes on which most people travel?

3. Examine how the support of the pattern <Berlin> <Leipzig> <Dresden> <Barcelona> <Madrid> <Salamanca> changes if you increase the max-gap from 1.1 to 2.1. Why is this the case?

4. Set the window size to 4.0 and the min support to 0.8. What can you say about the order in which all people visit all cities? What is the reason for this?

Analyzing DAX Trends

From the Yahoo! finance page you can get the historical data of a variety of stocks (http://finance.yahoo.com/q/hp?s=%5EGDAXI&a=02&b=01&c=1994&d=01&e=28&f=2014&g=d&z=66&y=5016). The dataset dax_94_14 which you can download from the website of the course, includes the stock numbers of the DAX from the 1th March 1994 till the 28th February 2014. In the following we will work with the Series Extension of the RapidMiner.

1.1 Visualize the trend of the DAX index (close) over the last 20 years. Experiment with different models (Regressions, SVM, Neural Net). What do you think reflects the general trend most? Use the Plot View with Series Plotter to visualize your results.

1.2. Now eliminate the randomness using the MOVING AVERAGE operator with a window of 30 and rebuild the trend on these values. Do the same with exponential smoothing.

2. Apply the Windowing operator for the Series Extension to your data set and set the window size to 5. Inspect the results to see what happens.

3. Now that we have an idea, how we can transform the data we can think about building a classifier to predict the labels of the data.

> Why does it not make sense to use Cross Validation?
>
> Build a classification process using “Sliding Window Validation”. Then test different values for the “horizon” parameter of the “Windowing” operator. How does your performance change?