Project Proposal: Will the Foreign Currency Exchange Rate Rise or Fall Tomorrow?

Xin Guo

March 13, 2014

1 Introduction

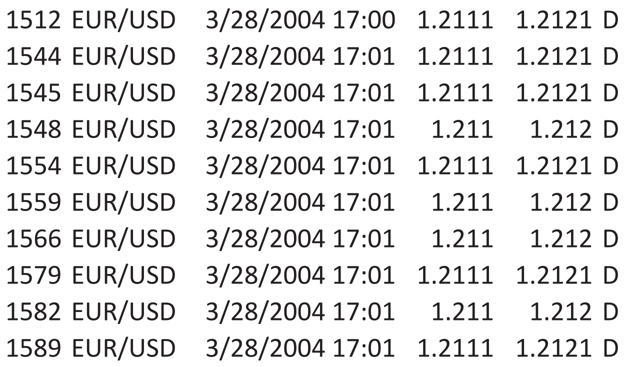
Prediction of the tendency in financial market is an old and important problem. A good prediction of the market (e.g. foreign currency exchange rate or stock price) may lead to large capital or prevent potential loss. The financial market is a complicate system and even a small part of the financial market is affected by many effects. For example, the stock price of a company can be affected by the development of the company or its competitor, the national economic situation, and the expectation of the market. The foreign exchange rate between two countries may be affected by the war in the third country which has important relationship with the two countries. The international gold price may be changed by the unexpected investigation from large group of individual investors. Due to its complicity, it is really challenging to make right predict by person.

As computer power develops, algorithms from fully-developed machine learning are implemented and successfully applied to many areas (e.g. criminal investigation, bioscience, health care). Previous research studies also showed that algorithms as back propagation neural networks (BPNN) can predict the stock price with relatively high accuracy. It is encourage to build up an accurate prediction model for financial market.

In this project, I plan to build up a model to predict the tendency of the foreign currency exchange rate, i.e., the rate will rise or fall on the next day, and find out the features, which are important to the prediction. With the limited time, I will focus on building up prediction models for the exchange rate between the currency pairs of the United States Dollar (USD) and Euro (EUR), USD and Great British Pound (GBP), and USD and Japanese Yen (JPY).

2 Data resource

The data is downloaded from `ratedata.gaincapital.com`. Figure 2 shows a snap shot of the data set. In the data set, the first column is trading ID, the second column the currency pair, the third column the trading time, the fourth column the bid rate, the fifth column the ask rate, and the six column the dealable indicator. Each row indicates the information of one tick. The information ranges from 2004 to 2014 for USD/EUR, USD/GBP, and USD/JPY. Each data file is a zip-compressed csv file. I plan to develop a web-scraper program to download all the data to local computer. The size of the raw data is expected to be 10 GB. Two days are needed for this step.



1512	EUR/USD	3/28/2004 17:00	1.2111	1.2121	D
1544	EUR/USD	3/28/2004 17:01	1.2111	1.2121	D
1545	EUR/USD	3/28/2004 17:01	1.2111	1.2121	D
1548	EUR/USD	3/28/2004 17:01	1.211	1.212	D
1554	EUR/USD	3/28/2004 17:01	1.2111	1.2121	D
1559	EUR/USD	3/28/2004 17:01	1.211	1.212	D
1566	EUR/USD	3/28/2004 17:01	1.211	1.212	D
1579	EUR/USD	3/28/2004 17:01	1.2111	1.2121	D
1582	EUR/USD	3/28/2004 17:01	1.211	1.212	D
1589	EUR/USD	3/28/2004 17:01	1.2111	1.2121	D

Figure 1: Snap shot of the raw data set.

3 Data preparation and analysis

I will use the data to build a model to predict the tendency of the foreign currency exchange rate. The output of the model will be that the highest/lowest exchange rate on the next day will rise or fall (i.e., 1 or -1). This is a supervised, classification problem. The detailed steps of analysis and model construction are listed as follows.

1. The first step is to reduce the dimension of the information space. For the exchange rate on each day, I will focus on the highest, lowest, mean, and closing rates and the standard deviation of the rate. I will develop a map/reduce program to find out the values of the five variables and put into a SQL database. One day is needed for this step.

2. The second step is to develop features of the data from the reduced information space. I will use the information obtained from the first step and develop Python program to calculate features for my prediction model. The total number of features will be about $100 \sim 150$. One day is needed for this step. The examples of features are shown as follows:

   (a) From data:

      i. Rates in previous 5 days (about 25 features).

      ii. Linear regression slope of rates in previous 2, 5, 10, and 30 days (about 20 features).

      iii. Taylor expansion based slope of rates based on previous 2, 5, 10, and 30 days (about 20 features).

   (b) From financial index

      i. Stochastic Oscillator (SO). A n-day-period SO is calculated as

$$SO = \frac{Current\ Close - Lowest\ Low}{Highest\ High - Lowest\ Low} * 100, \tag{1}$$

      where the highest high is the highest price in the n-day period, the lowest low is the lowest price in the n-day period, and current close is the closing price for today.

ii. Momentum.

$$Momentum = Current\ Closing\ Price - Oldest\ Closing\ Price, \tag{2}$$

where oldest closing price is the closing price of yesterday.

3. The third step is feature selection and extraction. I will use feature selection techniques, e.g., support vector machine recursive feature elimination (SVM-RFE), random forest, and feature extraction techniques, e.g., single value decomposition (SVD) to reduce the feature size to 5 to 20. Each subset of features from the feature selection/extraction algorithms will be used to build the model in the next step to find out the best combination of feature selection/extraction and classification algorithms. The major concern in this step is the computation time. Therefore, it is very important in designing the efficient algorithm for the computation. Two days are needed for this step.

4. The last step is to build up classification model. Based on the subset of features from the previous step, classification models, e.g., logistic regression, support vector machine (SVM), BPNN, radial basis function neural network (RBFNN), will be trained. The whole data set will be separated in to training data set (60 %), cross-validation data set (20 %), and test data set (20 %). The model will be fine tuned on the training and cross-validation data sets, and be tested on test data set. The same concern on computation time as the previous step applies here. Especially for BPNN, the training time may be long for the large data set. Possible solution is to choose small number of hidden layers and unites. Three days are needed for this step.

4 Final production

The final product will be a report/powerpint. The file should include summary of the model-building work flow, the comparison among the different combination of feature selection/extracton algorithms and classification models, the accuracy analysis of the model, and the prediction results of the model. If the time allows, a web-site with live time stream data can be built to predict the next day tendency with a report of the analysis.