Technical Documentation
TWFX

profor321@gmail.com
School of Computing and IT
University of Wollongong

June 3, 2017

R Programing Language

R programming language is used to write TWFX and the app is hosted using R Shiny which is a R web application framework. TWFX could accessed from TWFX subpage on Profor's website. R packages used in the software is listed below.

Packages
shiny
quantmod
plyr
dplyr
stringr
ggplot
RMySQL
lubridate
e1071
caret

Tweets Extraction

Tweets are extracted from 30 selected accounts and only 6 features are extracted from the original 17 features. The features that were extracted are text, favouriteCount, created, id, screenName, and retweetCount. text is the content of the tweet, favouriteCount is the frequency of that tweet being favourited by users while created is the date the tweet was created. id is the unique identifier of the tweet, screenName is the username of the twitter account of the tweet while retweetCount is the frequency of the tweet being retweeted by twitter users. Below is the list of accounts that tweets were extracted from. Extraction of tweets is performed using the R script named proforExtract.R.

Twitter Accounts
aus_business
barronsonline
BBCBusiness
Bloomberg
BloombergTV
BNN
bpolitics
business
BusinessDay
BusinessDesk
BW
CBCBusiness
ChinaBizWatch
EconBizFin
financialpost
FinancialReview
FinancialTimes
globebusiness
handelsblatt
LesEchos
markets
NAR
nikkei
ReutersBiz
SBH_USA
SkyNewsBiz
sole24ore
telebusiness
tijd
WSJ

Twitter Analysis

Sentiment for a foreign exchange is analysed by the sum of positive tweets and negative tweets. Positive tweets have positive effect while negative tweets have negative effect. Tweets are evaluated on the frequency of positive and negative words in the tweet. Positive and negative words were collected from `http://www.enchantedlearning.com/wordlist/positivewords.shtml` and `http://www.enchantedlearning.com/wordlist/negativewords.shtml`.

Data Set

Each sample is a day's trade in the foreign exchange market. The data set that is used to predict the trend of currency pair extracts two features. The first feature is the sum of the positive and negative tweets for the day of the first currency and is USD. The second feature is the sum of the positive and negative tweets of the second currency that is selected by the user. The labels for each sample is the upward outcome labelled Up or downward outcome labelled $Down$. The table below shows an example of how the data set could look like.

Att1	Att2	Label
13	8	Up
3	8	Down
1	7	Up

Support Vector Machines

Support Vector Machines (SVM) is used as the machine learning algorithm to model the classifier. Support vector machine is a maximal margin classifier for binary classification using hyperplane to linearly separate classes. A hyperplane is a flat affine subspace of dimension $p - 1$ in a p-dimensional space. In two dimensions, a hyperplane is defined by the equation $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ where any $X = (X_1, X_2)^T$ is a point on the plane. The optimal separating hyperplane that is the farthest from the training observations having the largest margin is used to linearly separate the two classes. In cases of non linear classification, it is addresed by enlarging the feature space using quadratic, cubic, and even higher-order polynomial functions of the predictors. SVM is also extended to classification of more than two classes using $one - versus - one$ and $one - versus - all$ approach. The $one - versus - one$ approach compares a pair of classes and classification is performed by assigning the sample to the class which it was most frequently assigned in the pairwise classification. The $one - versus - all$ approach fits K SVMs, each time comparing on of the K classess to the remianing $K - 1$ classes. An observation is assigned to the class with the highest level of confidence that the test observation belongs to the kth class. The SVM for the foreign exchange outcome classification uses C-classification with Gaussian kernel.

Best Parameters

The classification uses a grid search to find the best C and γ for SVM before building the model on SVM. The range used for the grid are $C = 2^{-5}, 2^{-3}...2^{15}$ and $\gamma = 2^{-15}, 2^{-13}...2^{3}$. The grid search is done using cross-validation with 10 fold. The best C and γ is then used to train the model with 90% training set and validated on 10% test set. The table below shows the results on USD against JPY on 5 experiments.

Parameters	Training Accuracy	Testing Accuracy
C = 8192, $\gamma = 2^{-15}$	50.66667%	50%
C = 2048, $\gamma = 2$	53.33333%	66.66667%
C = 32768, $\gamma = 50$	49.33333%	50%
C = 512, $\gamma = 2$	53%	66.66667%
C = 8192, $\gamma = 0.125$	53.66667%	66.66667%

Table 1: USD against JPY

Parameters	Training Accuracy	Testing Accuracy
C = 0.03125, $\gamma = 2^{-15}$	65.66667%	66.66667%
C = 0.03125, $\gamma = 2^{-15}$	66%	66.66667%
C = 128, $\gamma = 8$	68.66667%	100%
C = 8192, $\gamma = 0.5$	70%	83.33333%
C = 0.03125, $\gamma = 2^{-15}$	65.66667%	66.66667%

Table 2: USD against GBP

Conclusion

The models using SVM did not show very high accuracy but the results look promising. The tweets that were collected could only form less than 100 samples for each currency pair and this is not an ideal situation where machine learning algorithms need large data set to be optimal. No preprocessing was done for feature extraction and the results look promising. The favourite count and retweet count could be potential features but were not used as the tweet extraction script was not designed to update favourite count and retweet count. The extraction script was only designed to extract available tweets on the twitter's account timeline and favourite and retweet count of recent tweets usally do not reflect the sentiment of the tweets as it takes some time to build the favourite and retweet counts.