Statistical Arbitrage by Pair Trading using Clustering and Machine Learning.

QUANTITATIVE FINANCE AND RISK MANAGEMENT
UNIVERSITY OF MICHIGAN

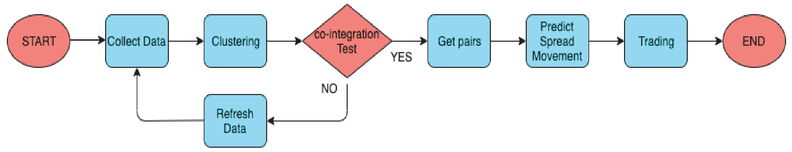
Authors: Israel Diego, Yingcong Jiang
Kaiqi Wang, Wuren Wang

Abstract

We investigate the application of machine learning methods to find statistical arbitrage opportunities in the stock market using pair trading strategy. Pairs are recognized using clustering methods, while trading signals are predicted by multiple supervised learning algorithms.

Motivation

The key to successful pairs trading is the ability to detect patterns in spreads and correctly identify when a spread is likely to converge back to its mean. Sophisticated machine learning techniques can be used at every step of the pairs trading process.



Methods

PCA: Feature dimension reduction and component generation, preparing for clustering

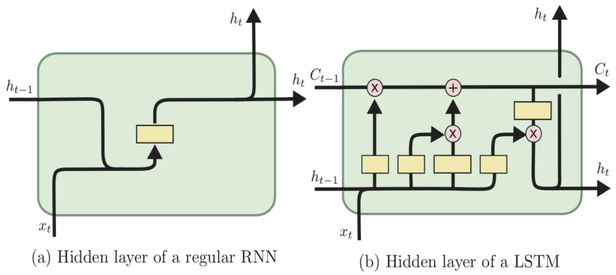
DBSCAN: Creates clusters and identify points that are not part of any cluster.

t-SNE: Method to visualize clusters from high dimension to 2-D space.

Gradient Boosting: A sequential ensemble model to capture complex patterns.

Random Forest: A bagging decision tree to reduce bias.

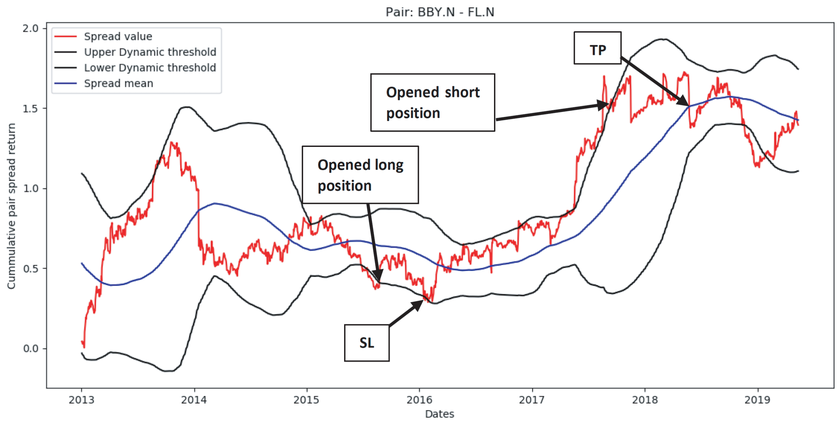
LSTM: Apply more weight to recent observations in time series prediction. Comparing with the standard RNN, LSTM diminishes the problems of long-term dependencies.

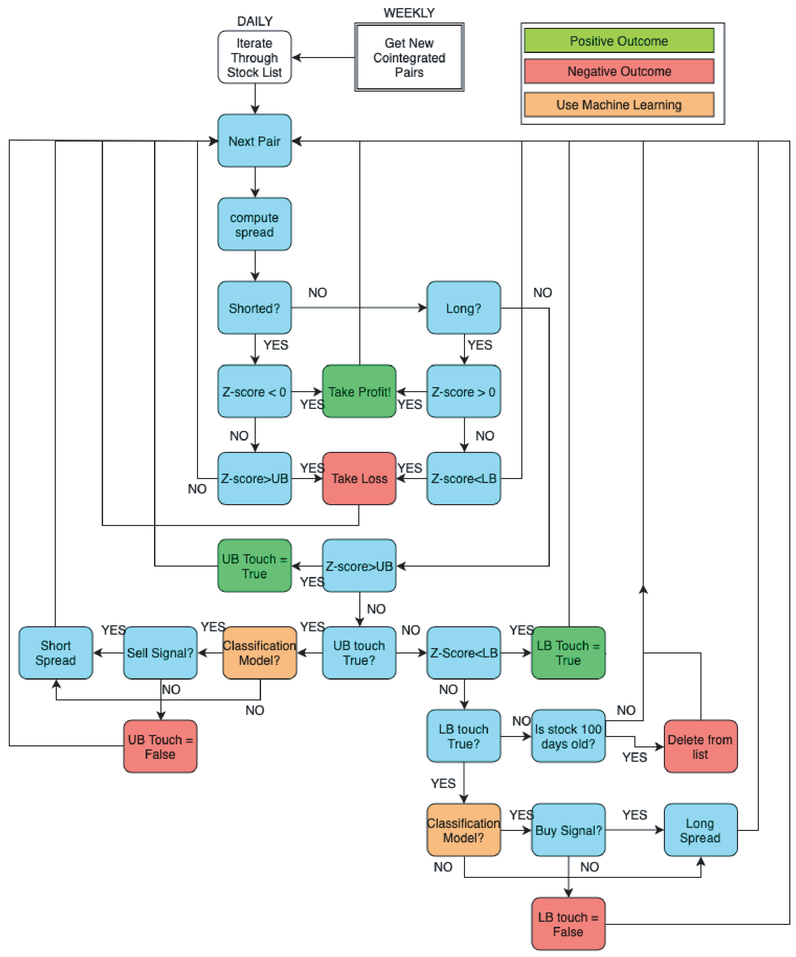


(a) Hidden layer of a regular RNN (b) Hidden layer of a LSTM

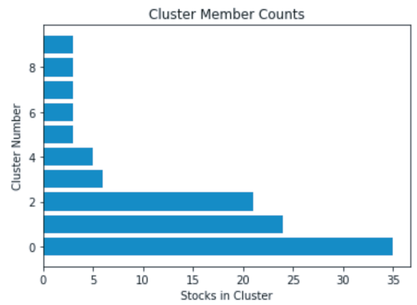
Model	NDAQ-US	ICE-US	CME-US
Baseline	0.502	0.506	0.513
Random Forest	0.63	0.67	0.58
Gradient Boosting	0.61	0.67	0.63
Logistic Regression	**0.69**	**0.76**	**0.70**
LSTM	0.56	0.53	0.61

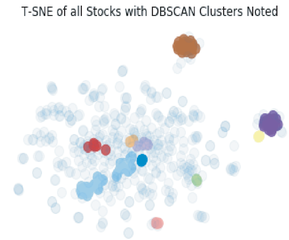
Trading Methodology:
Open long/buy position after the spread has hit the lower threshold from down-up. SL (Stop Loss) is triggered when the lower threshold is hit for the third time, before the mean was reached. Open short/sell position after the spread has hit the upper threshold from up-down. TP (take profit) is triggered when the mean spread is reached.





Clustering : Our approach to pairs trading is to apply PCA for dimension reduction on a large set of features in order to ease computation of DBSCAN filter for clustering stocks. Afterwards, we use cointegration test to extract all possible combinations of stocks in each cluster that are within 5% significance level.





Results

Metrics	Baseline	Alternative
Total Returns	-7.55%	10.62%
Sharpe Ratio	-0.86	1.07
Max Drawdown	-10.13%	-3.06%
Volatility	0.02	0.02

What we have learned

We gained experience on training neural networks, dimensionality reduction, and supervised learning on time-series data.

Conclusion

Pair trading is still a feasible trading strategy and machine learning can improve its performance.