Sequence-to-Sequence Learning of Financial Time Series in Algorithmic Trading

Philip Arvidsson philip@philiparvidsson.com

Tobias ̊Anhed anhedtobias@gmail.com

December 12, 2016

Abstract

\*PLACEHOLDER\* Time series forecasting—predicting future values of variables within a domain—is a largely unsolved problem in complex or chaotic domains such as weather (e.g. humidity, temperature or wind speed) and economics (e.g. currency exchange rates or stock prices). Long short-term memory (lstm) variants of recurrent neural networks (rnns) are powerful models that have recently been applied with success to predict future time series sequences from historical, sequential datasets; such as predicting the weather for the next twenty-four hours, or predicting the next sequence of words in natural language. In this paper, we will attempt to investigate the problem by applying lstm-based rnns, something that has not been done before. \*PLACEHOLDER\*

Keywords: time series, deep learning, lstm, algorithmic trading, machine learning

Contents

1 Introduction 2 1.1 Background and motivation . . . . . . . . . . . . . . . . . . . . . . . . . . 2 1.2 Current research . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3 1.3 Problem statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2 Methodology 4 2.1 Forskningstrategi . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4 2.2 Valda Tekniker . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4 2.3 Datainsamling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4 2.4 Dataanalysmetod . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

1

1 Introduction

Time series forecasting—predicting future values of variables within a domain—is a largely unsolved problem in complex or chaotic domains such as weather (e.g. humidity, temperature or wind speed) and economics (e.g. currency exchange rates or stock prices). Examining the problem with the latest models in the field of machine learning—lstm- based rnns—gives rise to new hope of deepening our understanding of the problem.

1.1 Background and motivation

Artificial neural networks (anns) have been applied for several decades to solve classifica- tion (determining what class a data point belongs to) and regression (determining the value of a dependent variable as a function of a data point’s coordinates) problems. The early versions of anns could only solve classification problems (Rosenblatt, 1958), while more recent ann models are also capable of handling regression problems (CITATION NEEDED). An ann, expressed in simple terms, is a virtual model of a biological brain, consisting of neurons, synapses and dendrites, modelled by the ann through cells with scalars, activation functions and weights.

The earlier models were unable to predict time series sequences (i.e. determining what class a data point belongs to with respect to its movement over time, or determining the value of a dependent variable as a function of the data point’s movement over time), something that was later solved through the introduction of rnns (Rumelhart et al., 1986), which can handle datasets with points that have a temporal component dictating their intradependence and distribution within the dataset, where each point in the sequence is correlated to previous (with respect to time) points through several known and unknown factors.

rnns have historically had problems with vanishing (or exploding) gradients, meaning they break down during training and become unable to model the problem well (Pascanu et al., 2012), but this problem was solved through the introduction of the lstm-based rnns (Hochreiter and Schmidhuber, 1997).

Regardless of whether the problem is approached as a matter of classification or regression, the temporal aspect renders algorithms not specifically designed for the task impotent since one single input dataset could belong to several different classes depending on where, in time or order, it appears in the data sequence. The problem should therefore

2

be examined with algorithms designed specifically for time series sequences, such as lstm-based rnns.

1.2 Current research

Currently, in the field of machine learning, hidden Markov models (hmms), dynamic Bayesian networks (dbns), tapped-delay neural networks (tdnns) and rnns are commonly used to handle sequential datasets, and have been applied with some degree of success to financial markets (Saad et al., 1998; Kita et al., 2012; Zhang, 2004).

Although much research is currently being done into the application of lstm-based rnns in financial markets, the research is still in its infancy, and although the lstm-based rnn has almost twenty years on its neck, it is just now that progress is being made to a significant degree. The volume of data created in the past couple of years—about 90% of all data ever created by humans (Devakunchari, 2014)—together with relatively cheap but very fast graphics processing units (gpus), have opened up the field of machine learning to more cost efficient research.

More recently, deep lstm-based rnns have been applied with some level of success in predicting future time series sequences from historical, sequential datasets; such as predicting the weather for the next twenty-four hours (Zaytar and Amrani, 2016), or predicting the next sequence of words in natural language (Sutskever et al., 2014).

1.3 Problem statement

Financial time series are used ubiquitously in algorithmic trading. In algorithmic trading, it is imperative that accurate predictions are made about numerous variables (such as volatility and returns) in order to time market entry and exit.

Since deep lstm-based rnns have only been applied within the algorithmic trading domain to a minimal extent, and since they have shown success in solving similar problems in other domains, it raises the question whether the technique can be used to predict a future sequence of financial variables that can be used to time both entry and exit positions within a certain time horizon.

Presuming that correlations exist along the temporal dimension of the dataset, the problem is reduced to a matter of finding an appropriate set of features enhancing the correlations, on which to train the lstm-based rnn. Expressed in a more concise manner, we attempt to answer the question:

Can financial markets be predicted accurately through the application of lstm-based rnns?

3

2 Methodology

Studien utfördes för att kunna analysera sv ̊arigheterna att förutsp ̊a framtida värden i en kaotisk marknad med dem senaste teknikerna för neurala nätverk. D ̊a problemet handlar om att tolka stora mängder data gjorde vi en kvantitativ ansats.

2.1 Forskningstrategi

Den kvantitativa ansatsen gjordes för att tolka all data över valutahandel som det neurala nätverket fick träna p ̊a. Vi studerade dess predicerade resultat för en valutas framtida riktning. Riktningen p ̊a en valutas värde kunde vara antingen uppg ̊ang, nedg ̊ang eller stillest ̊and.

2.2 Valda Tekniker

Det neurala nätverket implementerades i Python 2.7 med biblioteket TensorFlow (Abadi et al., 2015). Det neurala nätverket var ett rnn med lstm. B ̊ada teknikerna (rnn & lstm) finns implementerade i detta bibliotek. En avgränsning som behövde göras var just att använda TensorFlow. Det hade varit intressant att göra jämförelser mellan TensorFlow och andra stora bibliotek med implementeringar av neurala nätverk s ̊a som Theano, Keras, Lasagne, Blocks, SciKit m.m.. Att göra en egen implementering av ett neuralt nätverk för djupare konfigureringar fick även det prioriteras bort. Specifika hyperparametrar för TensorFlow ändrades under arbetets g ̊ang för att optimera nätverket. Olika storlekar p ̊a tidsfönstret testades för att hitta ett optimalt sätt för nätverket att förutsp ̊a valutan.

2.3 Datainsamling

Studien är avgränsad mot valutahandel d ̊a högfrekvens-data över aktier har en hög efterfr ̊agan. Det gör att den är sv ̊ar, men framförallt dyr, att f ̊a tag p ̊a. För att kunna genomföra studien behövdes data med hög frekvens samt att den har tillräckligt med parametrar utan att för den delen bli inkomplett. Datan som användes hämtades fr ̊an foreign exchange market (forex) och visar högfrekvenshandel med valutor, där fanns parametrarna som behövdes för att träna nätverket.

4

2.4 Dataanalysmetod

Resultaten som nätverket gav jämfördes sedan med korrekt data för samma tidsperiod och en slutsats kunde lätt dras ang ̊aende modellens pricksäckerhet. Det gjordes även jämförelser mot andra ”state-of-the-art”-algoritmer implementerade i just TensorFlow, detta för att ha möjlighet att utvärdera effektivititen av studiens nätverk.

5

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Devakunchari, R. (2014). Analysis on big data over the years. International Journal of

Scientific and Research Publications, 4(1).

Hochreiter, S. and Schmidhuber, J. (1997). Long short–term memory. Neural Computation,

9(8):1733–1780.

Kita, E., Zuo, Y., Harada, M., and Mizuno, T. (2012). Application of bayesian network

to stock price prediction.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). On the difficulty of training recurrent

neural networks.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage

and organization in the brain. Psychological Review, 65(6):386–408.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations

by back-propagating errors. Nature, 323:533–536.

Saad, E. W., Prokhorov, D. V., and II, D. C. W. (1998). Comparative study of stock

trend prediction using time delay, recurrent and probabilistic neural networks.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with

neural networks.

Zaytar, M. A. and Amrani, C. E. (2016). Sequence to sequence weather forecasting with

long. International Journal of Computer Applications, 143(11).

Zhang, Y. (2004). Prediction of financial time series with hidden markov models.

6

**1. Hur har redigeringen av introduktionen gjort, vilka förbättringar kan ni se i texten, och vilka avsnitt kommer att behöva ytterligare redigering innan slutlig inlämning.**

Vi ser att många förbättringar har gjorts i texten, finns källförteckningar där det saknades innan (bortsett från en, men den kommer väl xD). Det kändes som en tydlig “tratt”. Förklaringar på dom mest relevanta begreppen som saknades innan finns nu och det gör att man får en mycket tydligare förståelse för problemet.

**2. Fungerar problemställningen, är det en fråga som bör besvaras med en kvantitativ eller kvalitativ studie, eller är det ”valfritt”.**

Vi tror absolut att den med fördel bör besvaras med hjälp av en kvantitativ ansats.

**3. Hur väl passar metodförslag med problemställningen? Kommer författarna att få den information som de behöver genom de metoder de föreslår? Om inte, varför då?**

Ja det tror vi men utifrån att det är DSR med en kvantitativ ansats som ni valt. Det som kan vara bra är att identifiera empiri (den data ni faktiskt kommer att mäta på för att se om ni kan besvara forskningsfrågan) vilket jag tror i ert fall det resultat ni får utifrån evalueringen av den algoritmen ni valt med respektive parametrar. Sen en annan fråga som kan vara intressant är hur kan man (kan man överhuvudtaget?) styrka att de resultat som ni fått fram är bra, öka validiteten?

**4. Hur väl har författarna reflekterat över möjliga svagheter med de olika metodvalen? Vilken metod skulle ni själva välja om ni skulle göra samma studie?**

Saknar exempelvis jämförelse med andra metoder även reflektioner kring svagheter/styrkor. Skulle valt samma typ av studie.