Data article

Title: Stock Market Index Data and Indicators for Day Trading as a Binary Classification problem

Author: Renato Bruni

Affiliation: Dip. di Ingegneria Informatica, Automatica e Gestionale, Sapienza Università di Roma, Rome, Italy

Contact email: bruni@dis.uniroma1.it

Abstract

Classification is the attribution of labels to records according to a criterion automatically learned from a training set of labeled records. This task is needed in a huge number of practical application, and consequently it has been studied intensively and several classifications algorithms are today available. In finance, a stock market index is a measurement of the value of a section of the stock market. It is often used to describe the aggregate trend of a market. One basic financial issue would be forecasting this trend. Clearly, such a stochastic value is very difficult to predict. However, technical analysis is a security analysis methodology developed to forecast the direction of prices through the study of past market data. Day trading consists in buying and selling financial instruments within the same trading day. In this case, one interesting problem is the automatic individuation of favorable days for trading. We model this problem as a binary classification problem, and we provide datasets containing daily index values, the corresponding values of a selection of technical indicators, and the class label, which is 1 if the subsequent time period is favorable for day trading and 0 otherwise. These datasets can be used to test the behavior of different approaches in solving the day trading problem.

Specifications Table

Subject area	*Economics and Finance*
More specific subject area	*Financial Timing, Trading, Stock Index Analysis*
Type of data	Daily time series for two major indices belonging to two different stock markets
How data was acquired	*http://www.investing.com/*
Data format	*CSV format, can be opened as excel or text*
Experimental factors	*When necessary, data are filtered to remove missing or inaccurate data*
Experimental features	*The data provided consist of some series of daily prices (opening, closing, maximum, minimum), several series of technical indicators, and a class label. They can be used to apply day trading algorithms or classification algorithms*
Data accessibility	*Data is within this article*

Value of the data

- **The datasets are taken from real-world major financial markets and they are very recent: they range from 20th April 2010 to 12th July 2016**

- **The datasets contain a vast selection of financial indicators regarded as highly trend indicative by technical analysis**
- **The datasets are filtered and cleaned to remove data errors and missing**
- **These datasets can be used as benchmarks by researchers willing to test trading algorithms on real-world recent data**
- **These datasets can also be used as benchmarks to test classification strategies on publicly available difficult data**

Data

We provide daily time series for two major indices belonging to two different stock markets. The first one is the Standard & Poor's 500 (S&P 500), which is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ. This is one of the most commonly followed equity indices, and many consider it one of the best representations of the U.S. stock market. The second is the Financial Times Stock Exchange Milano Indice di Borsa (FTSE MIB), which is the primary benchmark Index for the Italian equity markets. It consists of the 40 most-traded stock classes on the exchange, and captures approximately 80% of the domestic market capitalization. For these indices, for each trading day ranging from 20th April 2010 to 12th July 2016, we provide the opening price, closing price, maximum, minimum, and a number of indicators regarded as highly trend indicative by technical analysis (see, e.g., [1,2,3] and references therein), as described in more detail in the next section. Each data record corresponds to one day.

We also provide a binary classification for each day: the class is 1 if the subsequent time period is favorable for day trading and 0 otherwise. Data are filtered to check and to correct missing or inaccurate data. Indicators which are computed using the n past observations are available only from the $(n+1)$-th record of the dataset. The class is not available for the last record. These missing data are encoded as 'N'. No other missing data appear in the dataset. Data cleaning is indeed an important issue for similar data (see, e.g., [4] for references on this widespread problem). The data provided can be used to test the effectiveness of technical analysis in predicting the trend, or to test the accuracy of classification algorithms.

Experimental Design, Materials and Methods

Each data record refers to one single trading day. Such a time period is indicated by a subscript $t \in \{1, \dots, m\}$. Each data record is identified with the date and it contains the following values:

- o_t denotes the opening price of the index;
- c_t denotes its closing price;
- max_t denotes its maximum price;
- min_t denotes its minimum price.

By using the above values, for each trading day we compute:

- the return $r_t = (c_t - c_{t-1})/c_{t-1}$ of the index;
- the percentage variation of the closing price $\delta_t = 100(c_t - c_{t-1})/c_{t-1}$.

After this, we compute the indicators described below. For each of them, the current value is denoted by a subscript t, the previous by $t-1$, etc.

Momentum

Momentum is conventionally regarded as the basic trend-following indicator. It shows trend by remaining positive while an uptrend is sustained, or negative while a downtrend is sustained. The momentum $M_t(n)$ of the current time period t is computed as the difference between the current closing price c_t and the closing price of n days ago c_{t-n}

$$M_t(n) = c_t - c_{t-n}$$

In our case we use n = 5.

Range: momentum can take any real value, either positive or negative. Positive values of momentum denote that the index trend is increasing, and vice versa.

EMA

Moving averages are widely used for the analysis of time series. A simple moving average (SMA) is the unweighted mean of the previous n data of the historical price data, most often the closing price. A weighted moving average (WMA) has multiplying factors to give different weights to the different prices. Usually, recent prices receive more importance than older prices. In particular, an exponential moving average (EMA) applies weighting factors which decrease exponentially in the past, however never reaching zero. $EMA_t\,(n)$ is computed using the current closing price c_i and the EMA of the previous day $EMA_{t-1}(n)$.

$$EMA_t\,(n) = \left(\frac{2}{\mathrm{n+1}}\right) c_t + \left(1 - \frac{2}{\mathrm{n+1}}\right) EMA_{t-1}\,(n) = \left(\frac{2}{\mathrm{n+1}}\right)\left(c_t - EMA_{t-1}\,(n)\right) + EMA_{t-1}\,(n)$$

In our case we use n = 12 and also n = 26.

Range: EMA has the same range of the price of the asset; in general it can take any real positive value.

MACD

Moving Average Convergence/Divergence (MACD) is an oscillator that should reveal changes in the strength, direction, momentum, and duration of a trend in a stock's price. The simplest version of MACD is the difference between two moving averages, one over a shorter period n and one over a longer period m .

$$MACD_t\,(n,m) = EMA_t\,(n) - EMA_t\,(m)$$

Further insight can be obtained by using a third moving average of the $MACD(n,m)$ itself over a period s, called "signal line" $SL(s)$. When $MACD(n,m)$ increases and crosses the signal line, it is a bullish signal; when it decreases and crosses the signal line, it is a bearish signal.

$$MACD_t\,(n,m,s) = (EMA_t\,(n) - EMA_t\,(m)) - SL_t\,(s)$$

In our case we use n = 12, m = 26 and s = 9.

Range: MACD can take any real value, either positive or negative. Positive values denotes that the index trend is increasing, and vice versa.

ROI

Return on Investment (ROI) is one way of considering profits in relation to capital invested. Usually it is the ratio between return and invested capital. In our case, we use the average return over the last n days, denoted by $\text{aver}\{r_t\ ,\ r_{t-1}\ , \dots , r_{t-n+1}\}$, and the current closing value.

$$ROI_t\ (n) = \frac{\text{aver}\{r_t\ ,\ r_{t-1}\ , \dots , r_{t-n+1}\}}{c_t}$$

In our case we use n = 10, 20 and 30.

Range: ROI can take any real value, either positive or negative. Positive values denote income, negative ones denote loss.

RSI

Relative Strength index (RSI) is a momentum oscillator that compares the magnitude of recent gains and losses over a specified time period to measure speed and change of price movements of a security. By defining the upward change as $u_t = c_t - c_{t-1}$ if $c_t > c_{t-1}$ and 0 otherwise, and the downward change as $d_t = c_{t-1} - c_t$ if $c_t < c_{t-1}$ and 0 otherwise, the relative strength $RS(n)$ is the average of the last n upward changes divided the average of the last n downward changes.

$$RS_t\ (n) = \frac{\text{aver}\{u_t\ ,\ u_{t-1}\ , \dots , u_{t-n+1}\}}{\text{aver}\{d_t\ ,\ d_{t-1}\ , \dots , d_{t-n+1}\}}$$

Then, RSI is computed as follows

$$RSI_t\ (n) = 100 - \frac{100}{1 + RS_t}$$

RSI is considered a signal of overbought when above 70 and a signal of oversold when below 30.

In our case we use n = 10, 14 and 30.

Range: RSI oscillates between 0 and 100. It is near to 0 when the corresponding upward changes are near to 0, it is near to 100 when the corresponding downward changes are near to 0.

STOCHRSI

Stochastic oscillators attempt to predict price turning points by comparing the closing price of a security to its price range. This concept can be applied to the RSI itself, obtaining the Stochastic RSI (SRSI). By computing the RSI range from its minimum in the last n periods $\min\{RSI_t\ ,\ RSI_{t-1}\ , \dots , RSI_{t-n+1}\}$ and its maximum in the last n periods $\max\{RSI_t\ ,\ RSI_{t-1}\ , \dots , RSI_{t-n+1}\}$, the SRSI is defined as follows.

$$SRSI_t\ (n) = \frac{RSI_t\ (n) - \min\{RSI_t\ ,\ RSI_{t-1}\ , \dots , RSI_{t-n+1}\}}{\max\{RSI_t\ ,\ RSI_{t-1}\ , \dots , RSI_{t-n+1}\} - \min\{RSI_t\ ,\ RSI_{t-1}\ , \dots , RSI_{t-n+1}\}}$$

In our case we use n = 10, 14 and 30.
Range: its range is between 0 and 1.

ATR

Average True Range (ATR) measures the degree of price volatility. The rage of a price is simply defined as $max_t - min_t$, the True Range (TR) extends it to yesterday's closing price if it was outside of today's range:

$$TR_t = \max\{max_t, c_{t-1}\} - \min\{min_t, c_{t-1}\}$$

Now, by denoting with $EMA_t(n, X)$ the exponential moving average of a generic X over the last n periods, we have that ATR is the exponential moving average of the TR:

$$ATR_t\ (n) = \ EMA_t(n, TR)$$

In our case we use n = 14.
Range: It is any positive value.

ADX

Average Directional Index (ADX) does not indicate trend direction or momentum, only trend strength. It is computed using the positive directional indicator (+DI), the negative directional indicator (-DI), and the Average True Range (ATR).

By defining the upmove as $up_t = max_t - max_{t-1}$ and the downmove as $dw_t = min_{t-1} - min_t$, if $up_t > dw_t$ and $up_t > 0$ then $+DM_t = up_t$, otherwise $+DM_t = 0$;
if $dw_t > up_t$ and $dw_t > 0$ $then$ $-DM_t = dw_t$, otherwise $-DM_t = 0$.
Now, recalling that $EMA_t(n, X)$ denotes the exponential moving average of X over the last n periods, we compute

$$+DI_t\ (n) = \frac{100\ EMA_t(n, +DM)}{ATR_t(n)}$$

$$-DI_t\ (n) = \frac{100\ EMA_t(n, -DM)}{ATR_t(n)}$$

ADX is finally computed as follows, with $\text{abs}(.)$ denoting the absolute value:

$$ADX_t\ (n) = 100\ \frac{EMA_t(n, \text{abs}\big(+DI - (-DI)\big)\ )}{+DI + (-DI)}$$

In our case we use n = 14.
Range: It ranges between 0 and 100. Generally, ADX values below 20 indicate trend weakness, and values above 40 indicate trend strength.

Williams %R

Williams %R is an oscillator that analyzes whether a stock or commodity market is trading near the high or the low, or somewhere in between, of its recent trading range.

$$\%R_t\ (n) = -100\ \frac{\max\{max_t\ ,\ max_{t-1}\ , \dots , max_{t-n+1}\} -\ c_t}{\max\{max_t\ ,\ max_{t-1}\ , \dots , max_{t-n+1}\} - \min\{min_t\ ,\ min_{t-1}\ , \dots , min_{t-n+1}\}}$$

In our case we use n = 14.
Range: It ranges between -100 and 0. A value of −100 means that the close today was the lowest low of the past n days, and 0 means that today's close was the highest high of the past n days.

CCI
Commodity Channel Index (CCI) is used to identify cyclical trends not only in commodities, but also equities and currencies. Define the Typical Price (TP) as follows.

$$TP_t\ = \frac{max_t + min_t + c_t}{3}$$

By computing the simple average over the last n periods of the typical price and its standard deviation, CCI is defined as follows.

$$CCI_t\ (\mathrm{n}) = \frac{TP_t - \mathrm{aver}\{TP_t\ ,\ TP_{t-1}\ , \dots , TP_{t-n+1}\}}{0.015\ \mathrm{dev}\{TP_t\ ,\ TP_{t-1}\ , \dots , TP_{t-n+1}\}}$$

In our case we use n = 20.
Range: The CCI fluctuates above and below zero. The constant 0.015 should ensure that approximately 70% - 80% of CCI values lay between −100 and +100.

UO
Ultimate Oscillator (UO) uses buying or selling "pressure", represented by where the daily closing price falls within the daily true range. The Buying Pressure (BP) and the True Range (TR) are computed as follows.

$$BP_t\ =\ c_t - \min\{min_t, c_{t-1}\}$$

$$TR_t = \max\{max_t, c_{t-1}\} - \min\{min_t, c_{t-1}\}$$

Then, the total buying pressure over the past n days is computed as follows.

$$AVG_t\ (n) = \frac{BP_t\ +\ BP_{t-1} + BP_{t-n+1}}{TR_t\ +\ TR_{t-1} + TR_{t-n+1}}$$

Such a total buying pressure is computed for short, intermediate and long time intervals, and the UO is:

$$UO_t\ (n, m, s) = 100\ \frac{4\ AVG_t\ (n) + 2\ AVG_t\ (m) + AVG_t\ (s)}{7}$$

In our case we use n = 7, m = 14 and s = 28.
Range: It ranges between 0 and 100.

Class

The problem of data classification is the attribution of labels to records according to a criterion automatically learned from a training set, that is a set of records that already have a class. Classification is a very important data mining task (see also [5]), and many classifications algorithms are today available (e.g., [6]). We assign the class to each record, so that any portion of the dataset can be used as training set. After this learning phase, the classification algorithm will be able to predict the class for the rest of the records. The accuracy of such a prediction can be computed by comparing it with the real class, which is the one given in the dataset.

The class that we assign to each daily record is 1 if the subsequent day is favorable for intra-day trading and 0 otherwise. Favorable for intra-day trading means that the increase between the opening price and the closing price of the same day is large enough for obtaining a profit by buying at the opening price and selling at the closing price. A threshold must be selected to define "large enough"; we select the value 0.3%, which should provide a reasonable opportunity for profit. Therefore, the class is defined as follows. Its prediction would allow to perform intra-day trading in the following day, as described above, or it could possibly be used to define inter-day trading strategies.

$$Class_t = \begin{cases} 1 & \text{if} \quad 100\,(c_{t+1} - o_{t+1})/o_{t+1} \geq 0.3 \\ 0 & \text{otherwise} \end{cases}$$

Note that the class of a given day is clearly not computable from the data available up to that day. However, we assigned it for the whole dataset by simply looking, for each day, at the following day.

According to technical analysis, there should be some kind of relation between the above described indicators at day t and the market evolution at day $t + 1$, that determines the class of day t (see for example [7]). The classification algorithm aims at discovering such a relation by predicting the class using the above described indicators. For an analysis of the existence of profit opportunities with respect to the market index, see also [8].

The datasets are provided in CSV format, that can be opened with MS Excel or as text file.

References

[1] Colby RW. The Encyclopedia of Technical Market Indicators (2nd edition). McGraw Hill, New York, 2003.

[2] Kirkpatrick CD, Dahlquist JR. Technical Analysis: The Complete Resource for Financial Market Technicians (3rd edition). Financial Times Press, Old Tappan, New Jersey, 2006.

[3] Murphy JJ. Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance, Paramus, New Jersey, 1999.

[4] Bruni R. Error Correction for Massive Data Sets. Optimization Methods and Software 2005 20:295-314.

[5] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.

[6] Bruni R, Bianchi G. Effective Classification using Binarization and Statistical Analysis. IEEE Transactions on Knowledge and Data Engineering 2015 27:2349-2361.

[7] Lo AW, Hasanhodzic J. The Evolution of Technical Analysis: Financial Prediction from Babylonian Tablets to Bloomberg Terminals. Bloomberg Press, New York, 2010.

[8] Bruni R, Cesarone F, Scozzari A, Tardella F. A Linear Risk-Return Model for Enhanced Indexation in Portfolio Optimization. Operations Research Spectrum 2015 37:735-759.