FORECASTING STOCK MARKET USING BAYESIAN NETWORKS: A TEXT MINING PROACH

Junwei Ma
Yanping Fu
Tiejun Wang
SWUFE
How Stock Price is Formed?

Stock Market  
Information  
Traders
Classification of Information

1. Digital information
   • Fama & French (1970)
   • Stiglitz (1980)
   • Kyle (1985)

2. Text information
   • Mitchell and Mulherin (1994)

3. Video information
   • ?
Stock market forecasting method

1. Traditional method
   - AR
   - MA
   - ARMA

2. Intelligence method
   - Natural networks, SVM, Bayesian Networks
Bayesian Network Forecast

• What’s Bayesian Network method?
  A Bayesian network, Bayes network, belief network, Bayesian model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.

• Why we use Bayesian Network method?
  1. Many factors, such as events, weather condition, can influence investors and then effect the stock price indirectly. So we try to use more other factors in Bayesian Networks to improve the prediction accuracy.
  2. Bayesian Networks can be used even some inputs are missed
Bayesian network with financial news information

The structure of our Bayesian Network model.

Compare with traditional time series models.

Rt-1, Rt-2...

Quantified financial information: stock message board.....

Rt
Why we choose “Gu Ba” but not other BBS?

1. According to a survey of the leading consulting group iResearch, EastMoney.com is the leader and the largest provider of financial information service in China.

2. We obtained 2,854,061 postings of 58 representative listed firms released from late 2007 to 30th March 2012. These postings are collected from the largest financial discussion board (EastMoney.com) in China.
Quantify Messages of firms on the discussion board

- A firm’s daily posts number from Jan.2009 to Mar.2012
Data Description

- Re-consider total postings according to text length, i.e. words count, 99% postings have no more than 86 words.
We divide all postings into five intervals, and count the postings in different intervals.

<table>
<thead>
<tr>
<th>Group</th>
<th>QUA1</th>
<th>QUA2</th>
<th>QUA3</th>
<th>QUA4</th>
<th>QUA5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>[0,23]</td>
<td>[24,43]</td>
<td>[44,63]</td>
<td>[64,85]</td>
<td>[86, max)</td>
</tr>
<tr>
<td>Percentage</td>
<td>0-30%</td>
<td>30%-50%</td>
<td>50%-70%</td>
<td>70%-99%</td>
<td>99%-100%</td>
</tr>
</tbody>
</table>
Relative Work

- Antweiler (2004)
- Mitchell (1994)

- Multiple Data Source, Multiple Statistic Method
  1. Text message is statistically significant to stock returns.
  2. The relationship is economically small

- In our work:
  1. If the relationship is really so slight?
  2. If our Bayesian network model should include this factor to forecast stock market?
Why returns but not price?

1. Is stock price should be concerned?

Situation 1.

<table>
<thead>
<tr>
<th>Period</th>
<th>Stock A</th>
<th>Stock B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td>$5</td>
<td>$6</td>
</tr>
<tr>
<td>Period 2</td>
<td>$6</td>
<td>$7</td>
</tr>
</tbody>
</table>

Which stock is dominating?

2. Compare with price series, return data is usually stable

Spurious return
Calculation of Idiosyncratic Return

• What’s idiosyncratic return?
  Part of an excess return on an asset not explainable by common factors, and independent of the specific returns of other assets.

• Why idiosyncratic returns?
  The correspondent text message from BBS is only against a special stock.

• How we calculate idiosyncratic returns?
  stock returns, intercept term, market excess returns, firm size, book to market value, idiosyncratic return. (Fama three factor)

\[
\bar{r}_t^i = \alpha_t^i + \beta_{MKT}^i \times MKT_t + \beta_{SMB}^i \times SMB + \beta_{HML}^i \times HML_t + r_t^i
\]
Regression with Total New Post Numbers

<table>
<thead>
<tr>
<th>Idiosyncratic Return</th>
<th>Number of daily Postings</th>
<th>Leverage Effect</th>
<th>C</th>
<th>Trading Volume</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.21E-05</td>
<td>--</td>
<td>-1.385E-03</td>
<td>5.93E-05</td>
<td>0.0209</td>
</tr>
<tr>
<td>(Std.Error)</td>
<td>(3.69E-07)</td>
<td>--</td>
<td>(3.36E-05)</td>
<td>(5.33E-07)</td>
<td></td>
</tr>
<tr>
<td>(t Value)</td>
<td>(-32.72)</td>
<td>--</td>
<td>(-41.19)</td>
<td>(-111.41)</td>
<td></td>
</tr>
<tr>
<td>Volatility</td>
<td>2.09E-06</td>
<td>-2.66E-04</td>
<td>6.14E-05</td>
<td>1.98E-06</td>
<td>0.0767</td>
</tr>
<tr>
<td>(Std.Error)</td>
<td>(1.80E-08)</td>
<td>(2.46E-06)</td>
<td>(1.96E-06)</td>
<td>(2.61E-08)</td>
<td></td>
</tr>
<tr>
<td>(t Value)</td>
<td>(116.40)</td>
<td>(108.15)</td>
<td>(31.37)</td>
<td>(76.11)</td>
<td></td>
</tr>
</tbody>
</table>

Has the similar result, so BBS is an reliable data source to represent public opinion.
<table>
<thead>
<tr>
<th>Idiosyncratic Return</th>
<th>C</th>
<th>QUA1</th>
<th>QUA2</th>
<th>QUA3</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficents</td>
<td>-1.40E-03</td>
<td>6.94E-05</td>
<td>-2.00E-04</td>
<td>-7.10E-05</td>
<td></td>
</tr>
<tr>
<td>Std. Error</td>
<td>-3.36E-05</td>
<td>-1.68E-06</td>
<td>-3.79E-06</td>
<td>-7.68E-06</td>
<td></td>
</tr>
<tr>
<td>t-value</td>
<td>-41.6358</td>
<td>41.37513</td>
<td>-52.82047</td>
<td>-9.248943</td>
<td>0.077</td>
</tr>
<tr>
<td>coefficents</td>
<td>7.83E-05</td>
<td>1.21E-04</td>
<td>6.03E-05</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Std. Error</td>
<td>-1.40E-05</td>
<td>-3.18E-06</td>
<td>-5.32E-07</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>t-value</td>
<td>5.604786</td>
<td>38.11191</td>
<td>113.4495</td>
<td>--</td>
<td></td>
</tr>
</tbody>
</table>
### Regressions with New Postings in Different Words Intervals

<table>
<thead>
<tr>
<th>Volatility</th>
<th>C</th>
<th>QUA1</th>
<th>QUA2</th>
<th>QUA3</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficients</td>
<td>7.41E-05</td>
<td>8.57E-06</td>
<td>-3.60E-06</td>
<td>-4.99E-07</td>
<td>0.088</td>
</tr>
<tr>
<td>Std. Error</td>
<td>-1.95E-06</td>
<td>-8.16E-08</td>
<td>-1.85E-07</td>
<td>-3.74E-07</td>
<td></td>
</tr>
<tr>
<td>t-value</td>
<td>28.01478</td>
<td>104.9874</td>
<td>-19.46463</td>
<td>-1.335186</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QUA4</th>
<th>QUA5</th>
<th>VOL</th>
<th>LEVER</th>
</tr>
</thead>
<tbody>
<tr>
<td>-7.14E-06</td>
<td>-1.79E-06</td>
<td>1.91E-06</td>
<td>2.62E-04</td>
</tr>
<tr>
<td>Std. Error</td>
<td>-6.80E-07</td>
<td>-1.55E-07</td>
<td>-2.60E-08</td>
</tr>
<tr>
<td>t-value</td>
<td>-10.50374</td>
<td>-11.54368</td>
<td>73.50909</td>
</tr>
</tbody>
</table>
• Conclusion of this regression:

The correlation between different words interval groups are different even opposite, this can explain why statistic result in other papers is significant but slight.

Thus the words of new post is an important factor, it can be used in the next Bayesian network forecasting.
Future works

• We will quantify other text information on the discussion board, which might be related to stock price.
• Introduce more other factors (variables) into Bayesian Networks and try to check whether this can improve the prediction accuracy.
Thanks