

Knowledge Discovery in Stock Market Data

Alfred Ultsch and Hermann Locarek-Junge

Abstract This work presents the results of a Data Mining and Knowledge Discovery approach on data from the stock markets using Databionics techniques. Stock market data is analyzed using methods that were learned from nature and previously applied primarily to DNA microarray data. It is demonstrated that the discovery of new insights into the stock markets is possible by the application of sensible preprocessing of daily returns (Relative Differences), application of a projection which has the potential to show emergent structures in the data (U-Matrix) and allows for a nontrivial clustering of the data (U*C).

1 Introduction

An issue that is the subject of intense debate among academics and financial professionals is the Efficient Market Hypothesis (EMH). It states that security prices fully reflect all available information at any time. The implications of the EMH are truly profound. Most individuals that buy and sell stocks in practice however, do so under the assumption that the securities they are buying are worth more than the price that they are paying, while securities that they are selling are worth less than the selling price.

Empirical evidence has been mixed, but has generally not supported strong forms of the efficient markets hypothesis, e.g. low P/E stocks have greater returns. Earlier papers also refuted the assertion that higher returns could be attributed to higher beta, which has been accepted by efficient market theorists as explaining the anomaly in neat accordance with modern portfolio theory. One can also identify “losers” as stocks that have had poor returns over some number of past years. “Winners” would be those stocks that had high returns over a similar period. Some trading rules say that in trends one should buy “winners” and sell “losers”. While proponents of the EMH don’t believe that it is possible to beat the market, some

A. Ultsch (✉)
Databionics Research Group, University of Marburg, Germany
e-mail: ultsch@informatik.uni-marburg.de

believe that stocks can be divided into categories based on risk factors. However, these risk factors are considered to be stable over time. In this paper, we analyze a very large stock market to find out whether there exist groups of stocks and clusters of time, where the groups that we find behave similar in the way that the probability of rising or falling stock prices within the created groups can be forecasted and is different from randomness, which would challenge the EMH.

2 Daily Returns on Stocks

Primary data in this paper are the adjusted daily closing prices of stocks traded in the USA. The prices of 7031 stocks were collected from Yahoo Finance (finance.yahoo.com) for the period Jan. 1st 2000 to 1st of March 2008 (observation period). This resulted in 2047 trading days. A total of 14,390,410 stock prices were obtained in this way. Standard & Poor's 500 Index – S&P 500 gives an overall picture of the market situation during the observation period (see Fig. 1). The S&P 500 is one of the most commonly used benchmarks for the overall U.S. stock market. It can be seen that the observation period rising as well as falling market conditions.

For each day (t) and each price $p(t)$ the daily return was calculated as Relative Difference ($RelDiff(t)$):

$$RelDiff(t) = 2 * (p(t) - p(t - 1)) / (p(t) + p(t - 1))$$

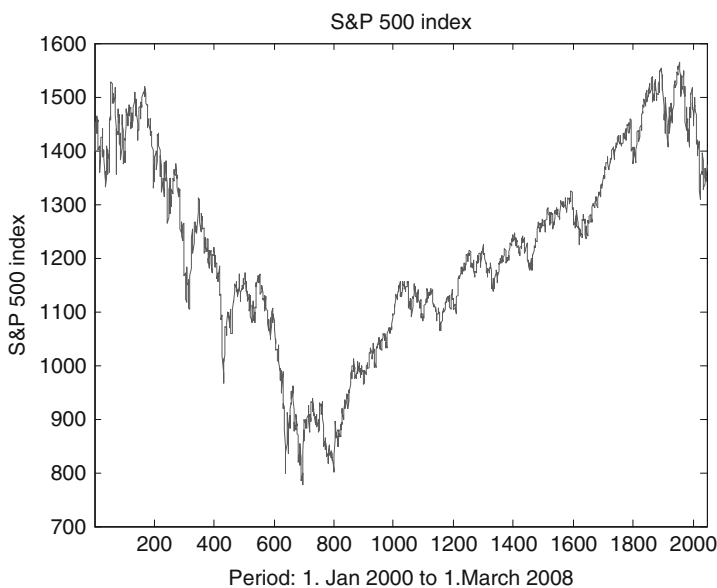


Fig. 1 S&P 500 during observation period

Relative Difference has several advantages over other formulas for return, like LogRatio ($\log(p(t)/p(t - 1))$) or Ratio $(p(t) - p(t - 1))/p(t - 1)$. See [Ultsch \(2009\)](#) for a detailed discussion. The most important for this investigation is that RelDiff possesses a symmetric and finite range: if a company defaults ($p(t) = 0$) then $RelDiff = -200%$. If a company has exorbitant gains ($p(t - 1) < p(t)$) then RelDiff approaches $+200%$. This allows to model returns with finite variances. In [Ultsch \(2009\)](#) it was shown that returns measured in RelDiff can be modeled with a mixture of distributions using one Normal (Gaussian) and two LogNormal distributions. The definition of logarithms was generalized to negative numbers as $\log'(x) = \text{sign}(x) \cdot \log(\text{abs}(x))$. An initial LogNormal, Gaussian, LogNormal (LGL) model was fitted to the data using the Expectation Maximization algorithm (e.g. [Izenman 2008](#)). Figure 2 shows the empirical probability distribution measured with a kernel density estimator Pareto Density Estimation (PDE) ([Ultsch 2003](#)). The LGL model is depicted in Fig. 2 using dashed lines for each component and a solid line for the mixture. The quality of the fit was assessed with a quantile/quantile plot resulting in an extremely good fit (see [Ultsch 2009, Fig. 5](#)).

This model can be naturally interpreted as a random result for returns, i.e. the central Gaussian $N(0, 1.7)$ with a fraction of 75% of all returns. Furthermore there are two non random distributions for returns, losses (12.5%) and wins (12.5%), which are lognormal distributed.

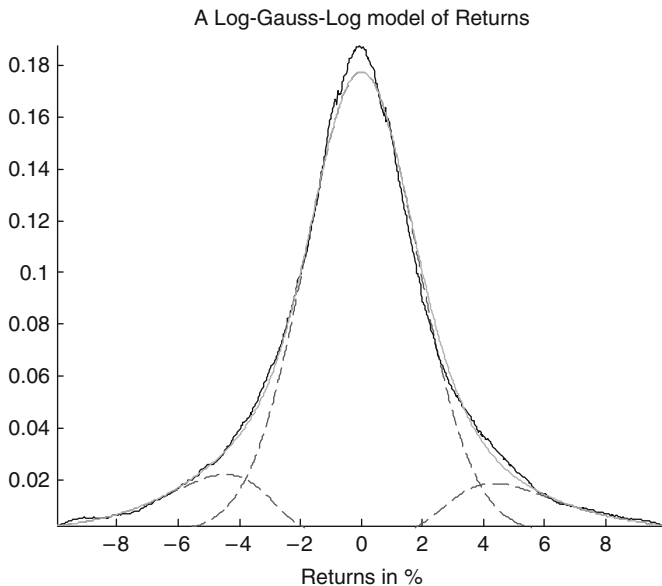


Fig. 2 The Log-Gauss-Log model of all returns

3 Knowledge Discovery in Market Activities

Using the model described in the last chapter it can be decided, whether a return belongs to the Random, Losses or Wins class using Bayes decision. We define $UnitWin = p(Return > Random) - p(Return < Random)$, where the probabilities are calculated with Bayes' theorem on the model developed above. UnitWin gives -1 for Losses, 0 for Random and $+1$ for Wins. In Fig. 3 UnitWin is shown for all returns.

The advantage of UnitWin is that differences in returns within same group are zero. UnitWin is therefore a good measure to compare the performance of different stocks for all trading days. The market activity on each day can be measured as the average number of non random returns for that day. This gives

$$Activity(t) = \underset{i}{mean}(abs(UnitWin(t, i)))$$

The distribution of Activity is shown in Fig. 4 using PDE.

It can be seen that Activity can be modeled as a mixture of Gaussians GMM (see Fig. 4). Using this GMM active days and inactive days can be distinguished. We found that the market was active for 2,045 days during our observation period. The next question is, whether there are days with more than average performance of the stock market. We defined the DailyPerformance(i) of a as the sum of all UnitWins for stock i . We found that the DailyPerformance consisted of three different distributions: a Gaussian around zero, i.e. passive performance or sideways

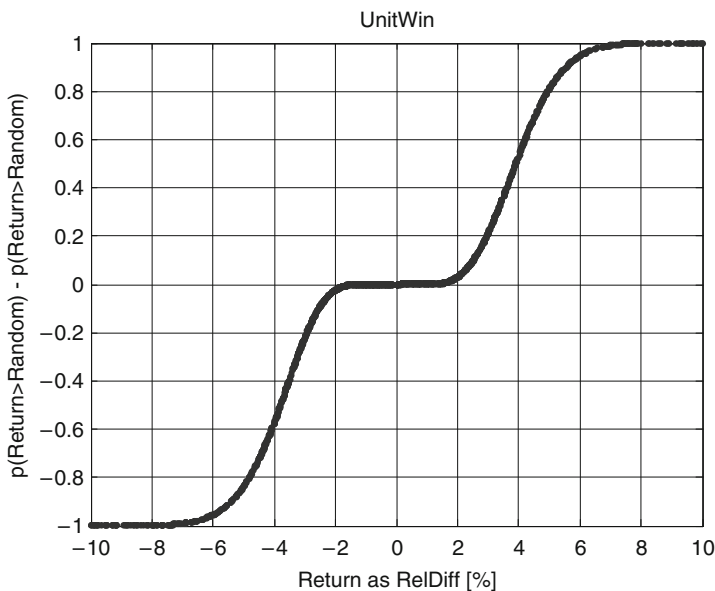


Fig. 3 UnitWin as a function of stock's returns

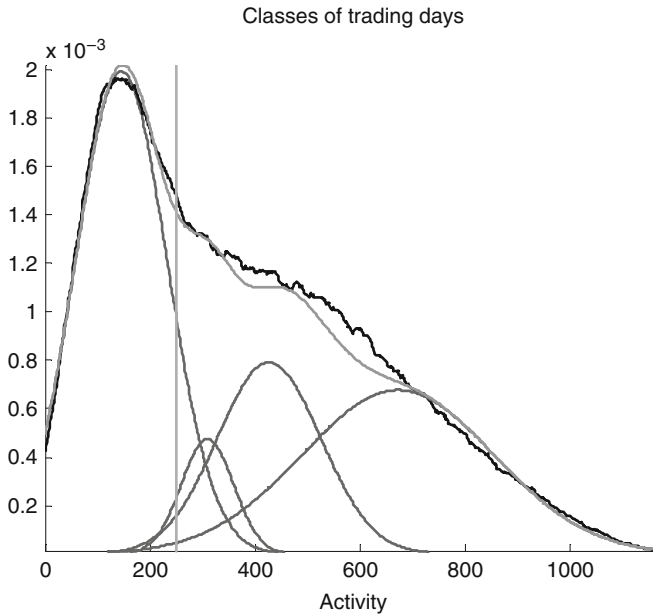


Fig. 4 Distribution of market activity for all days

movement of stocks and a winner and loser distribution, both lognormal distributed. Furthermore we found that only $568 = 8\%$ of all stocks dominated the performance of the stock market (marked leaders).

4 Types of Marked States

With UnitWins returns can be compared for sets of stocks and groups of days. The similarity respectively dissimilarity of marked days was defined as the Euclidean distance of the UnitWins of a set of stocks. Using this distance definition the different types of market days (winning, losing and passive) were compared for the marked leaders. For each of these groups a clustering procedure was performed using Emergent Self Organizing Feature Maps (ESOM) with the U-Matrix display (Deboeck and Ultsch 2000) and the clustering algorithm U*C (Ultsch 2007). Figure 5 shows an example of a U-matrix.

This 3D landscape is interpreted as follows: data in valleys are close in the high dimensional input space. Data separated by mountains are in different clusters. The $U * C$ clustering resulted in three clusters for Winner days (w_1, \dots, w_3), four classes for Loser days (l_1, \dots, l_4) and only one class for the Passive days.

As a next step the transition frequencies for each class were counted. The results is shown in Fig. 6. It is remarkable, that some states are rather persistent. For class 13, one of the loser classes, the probability that the next day is also a loser class is 74%.

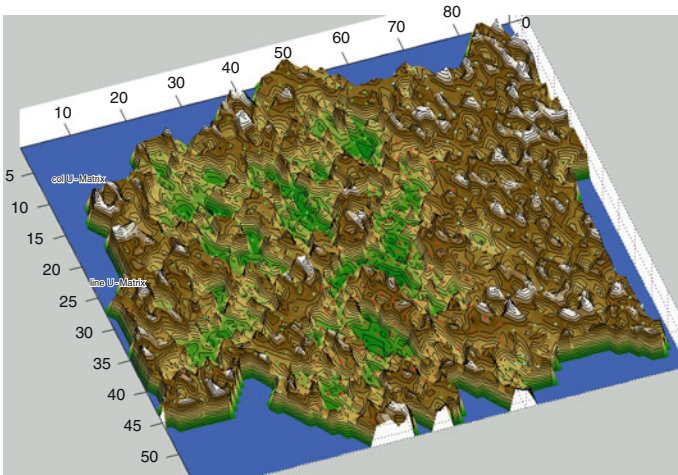


Fig. 5 U-Matrix of the winner days for the market leaders

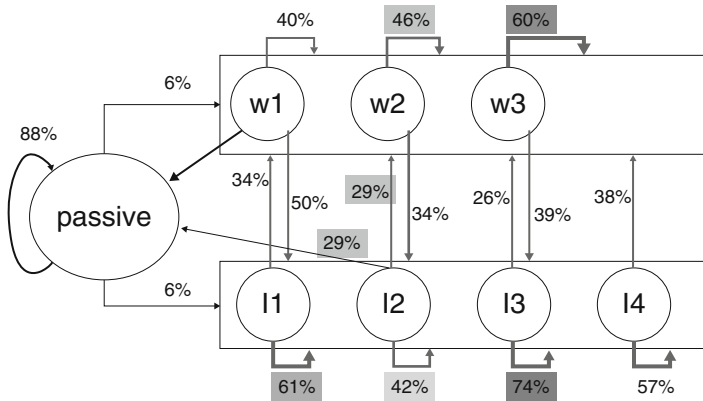


Fig. 6 Transition frequencies for the market classes

For class w_3 it was also observed that with a probability of 60% the next day is also a winning day. Other states are instable. E.g. in loser state l_2 with 58% probability, the next day is either passive or winning.

5 Discussion

This paper is an example of knowledge discovery in stock marked data. Knowledge Discovery is defined as the discovery of understandable knowledge which is new and useful. We have found that there are three types of returns: random, losses

and wins. Using Bayesian decision a meaningful aggregation of collective behavior could be defined (UnitWin).

Market activity was found to be either active or inactive. Daily performance could be classified in passive, winning and losing. UnitWin can be used for the definition of a sensible distance function. It has the advantage that inner group differences, e.g. within passive stocks or days, are zero.

Inter group differences contribute a precise and finite value to the distance function. Using this distance function, a clustering of the winner and loser market days was possible. The usefulness of these clusters can be seen in the transition frequencies to other. Some of the states suggest the buying (e.g. l_2) others the selling of stocks (e.g. w_1). It was not intended that this works may be used for the generation of buy-or-sell signals. It may, however, be useful to calculate measures for the overall state of a market day.

6 Conclusion

The EMH is the backbone of classical capital market theory. It has been tested empirically quite often, using econometrical testing and event studies. Several anomalies have been found, but they could mostly explained by applying risk measures and models for investor utility.

In this paper, knowledge discovery in stock marked data is applied. In the paper we found that there are three types of returns: random, losses and wins. A meaningful aggregation of collective behavior was defined and market activity was found to be either active or inactive while performance could be classified in passive, winning and losing. A clustering of the winner and loser market days was possible, where some of the states suggest buying, others the selling of stocks. It was not intended that this work may be used for the generation of buy-signals or sell-signals. It may, however, be useful to calculate measures for the overall state of a market day.

The authors will try to test the properties out-of-sample and in various other markets to find out whether the method works only in the sample period or it is a general property of the stock market, which remains to be proven with an independent test set. This challenge for the EMH remains future work.

References

- Deboeck, G. J., & Ultsch, A. (2000). *Picking stocks with emergent self-organizing value maps* (Vol. 10, pp. 203–216). Prague: Neural Networks World, Institute of Computer Science.
- Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Berlin: Springer.
- Ultsch, A. (2003). *Pareto density estimation: A density estimation for knowledge discovery*. D. Baier and K. D. Wernecke (Eds.), *Innovations in classification, data science, and information*

- systems – *Proceedings 27th annual conference of the german classification society (GfKI)* (pp. 91–100). Berlin, Heidelberg: Springer.
- Ultsch, A. (2007). Analysis and practical results of U*C clustering. *Proceedings 30th annual conference of the german classification society (GfKI 2006)*. Berlin, Germany.
- Ultsch, A. (2009). Is log ratio a good value for measuring return in stock investments? In: A. Fink, B. Lausen, W. Seidel & A. Ultsch (Eds.): *Advances in Data Analysis, Data Handling and Business Intelligence*, (pp. 505–511). Springer.