Stock Sector Analysis through Data Mining

Akira Madono, Amin Y. Teymorian

CSCI 243 - Data Mining
The George Washington University

akiramadono@gmail.com,amin@gwu.edu

Abstract—
We investigate the suitability of a typical industry-specific stock classification (the Global Industry Classification Standard) for representing stocks with similar price movements. In particular, we cluster high dimensional S&P 500 stock time series over a 7 month period using Symbolic Aggregate Approximation (SAX). Our results offer evidence that suggest that typical industry-specific stock sectors fail to adequately capture similar price behavior.

I. Introduction

Industry sector classifications, such as services, technology, or utilities, are often employed in asset allocation decisions [1]. The intuition is that stocks in the same sector react similarly to market conditions. For example, technology sector stock prices are less affected by the cost of oil than transportation sector stock prices. A portfolio that invests in stocks from both sectors can offset its exposure to market downturn in one of the sectors and hence reduce potential losses. We want to measure the extent to which sector classifications relate to price movement; how *suitably* stock sectors describe price movement.

II. Related Work

Applications of data mining to problems in finance have been of consistent interest over the past few years [2–5]. For example, in [2], artificial neural network formulations are employed to predict future stock prices over short-term horizons (e.g., 10 days). Although experimental results demonstrate good performance, successful short term predictions may be more attributable to "momentum" and luck rather than skill. Ref. [5] recently proposed a method for portfolio construction of stocks predicted to be top-performers.

Additionally, clustering stocks according to risk-return ratios has been proposed in [3]. The idea is to cluster stocks with a similar risk-return criterion, and invest in stocks with less geographic risk from the cluster of interest.

Most closely related to our work is the study presented in [4]. Rosen hierarchically clustered a small dataset, using correlation as his distance metric and provided a reasonable sector based rationale for that clustering. While we also find reasonable clusterings based on sector characteristics, we find the more often than not, sectors are not useful predictors of price behavior. Second, our data set is much larger although shorter in time. Third, we use SAX instead of correlation.

III. Method

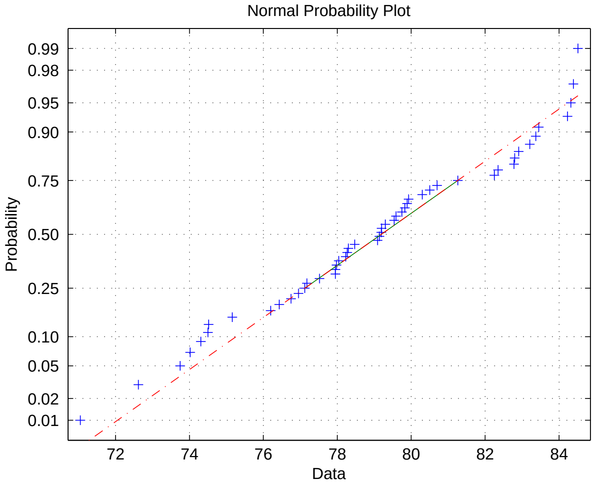
The Sample - On April 16, 2008, We defined a time period of 7 months and read S&P 500 stock closing prices from the yahoo finance website. Two companies, Phillip Morris International (NYSE:PM), and Teradata Corporation (NYSE:TDC) could not be included because they are spinoff corporation created within our time period. With a combined market capitalization of appx. 1%, we don't feel their exclusion will significantly detract from our findings.

Data Approximation - In order to approximate our data, we needed to control for rapid fluctuations and high dimensionality. We decided to use the Symbolic Aggregate Approximation (SAX) method proposed by E. Keogh of UC Riverside [6]. SAX is a refinement of Piecewise Aggregate Approximation (PAA), which splits up a given time series into a series of equally spaced time segments and assigns the mean value of each segment to its approximation of the time series. Formally, given a series $S = \{S_1, \ldots, S_n\}$, we create a series $P = \{P_1, \ldots, P_w\}$, where $w < n$ and any given element of P_i, is:

$$P_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} S_j \qquad (1)$$

Clearly, PAA reduces the dimensionality of the series.

Another one of its features is that it lower bounds the Euclidean Distance. A proof is given in [7]. SAX goes further by discretizing each PAA estimate along a Gaussian distribution for values taken from a moving window along the time series. SAX can thus both capture local activity while enabling dimensionality reduction. We refer the interested reader to [6]. The refinement of SAX over PAA is first allowable because SAX lower bounds PAA, and by transitivity, the Euclidean. Further, it is advantageous because it is able to generate an equiprobable discretized representation; an added benefit that enables use of popular Bioinformatics techniques. All that it requires is that your data be normally distributed within the space of a moving window. An established notion for stocks; the common assumption in asset pricing (a. la. Louis Bachelior circa. 1900) is that within an year's worth of periodic activity, stocks are normally distributed and lognormal more generally. For example. Here is the normal probability plot for 3M, for the 50 business days following September 04, 2007:



Other time periods, greater than 20 days followed a similarly normal behavior. Safe to say, our choice of 40 days, roughly two months for stock activity, safely fits within the normality assumption. Another parameter for SAX, is the alphabet size. E. Keogh discovered diminishing returns from alphabet sizes of 8 forward for a number of datasets. We chose the alphabet size of 8, in light of his empirical findings.

The Distance Metric: Generating the Distance Matrix - Given the above symbolic approximation, the next step in our process was to generate a distance matrix, to gage the similarity of our various stocks. Given the PAA reduction given in (1), we prepare the series for SAX by discretizing the PAA series along a Gaussian distribution divided into equal area segments. Let $\hat{Q}$ and $\hat{C}$ be two such discretized series. The distance metric that lower bounds SAX is $mindist$ [7] where.

$$mindist(\hat{Q}, \hat{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (dist(\hat{q}_i, \hat{c}_i)^2)} \qquad (2)$$

$dist()$ function is given by the following.

Let a be the alphabet count and $\beta = \{\beta_1 \dots \beta_{a-1}\}$ such that the area under a $N(0,1)$ Gaussian curve from β_i to $\beta_{i+1} = 1/a$. Let β_0 and β_a be defined as $-\infty$ and $+\infty$ resp. The distance between two alphabet values a_i and a_j is:

$$dist(a_i, a_j) = \begin{Bmatrix} 0 \text{ if } |i-j| \leq 1 \\ \beta_{max(a_i,a_j)-1} - \beta_{min(a_i,a_j)}, \text{ otherwise} \end{Bmatrix} \qquad (3)$$

Here is a sample mindist table for an alphabet of size 8:

TABLE I

MINDIST TABLE

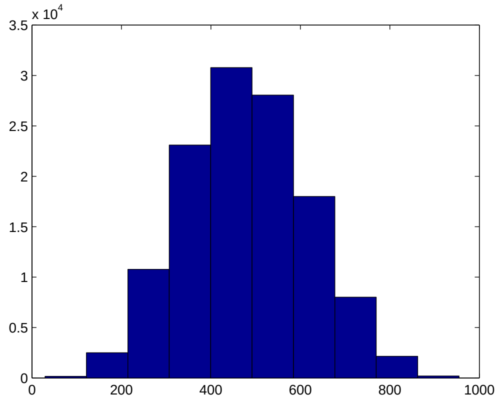
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0	0	0.48	0.83	1.15	1.47	1.82	2.3
2	0	0	0	0.35	0.67	0.99	1.34	1.82
3	0.48	0	0	0	0.32	0.64	0.99	1.47
4	0.83	0.35	0	0	0	0.32	0.67	1.15
5	1.15	0.67	0.32	0	0	0	0.35	0.83
6	1.47	0.99	0.64	0.32	0	0	0	0.48
7	1.82	1.34	0.99	0.67	0.35	0	0	0
8	2.3	1.82	1.47	1.15	0.83	0.48	0	0

We refer the reader to [7] for a more thorough treatment of the $mindist$ measure as we feel it is important but best described by its creators. Again, we have chosen it for its proven lower-bounding properties. Generating the distance matrix is a simple matter evaluating the $mindist$ function along a compressed 40 day window, and amongst all pertinent combinations of our 498 stocks.

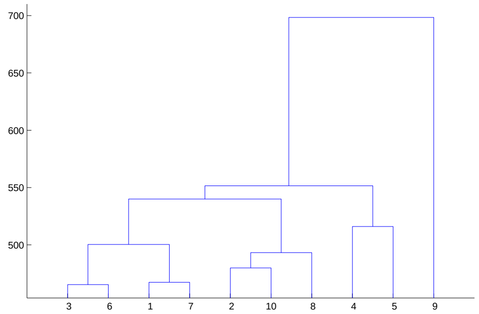
Hierarchically Clustering the Time Series - Large-cap stock data is generally highly correlated because all are affected by market forces; the activities of most companies are predicated on those of others including those in other lines of business. Hierarchical clustering best captures this interdependent nature.

We selected the unweighted average distance algorithm, easily implemented in MATLAB 7.1, because of the uniformly dense nature of our data. Given that our S&P 498 were originally classified into 10 sectors, we felt the best comparison would similarly classify into 10 sectors.

IV. RESULTS



First, we produced a distance matrix. In the figure above, we see the range of distance values produced from SAX. A much larger range, than starting value testifies to its effectiveness as an approximation method.



The above is the 10 cluster dendrogram produced by the unweighted average distance algorithm. Each of the leave numbers corresponds to a cluster. The cluster sizes in order of the above dendrogram are, $|C_3| = 14, |C_6| = 9, |C_1| = 249, |C_7| = 16, |C_2| = 96, |C_10| = 75, |C_8| = 16, |C_4| = 18, |C_5| = 4$, and $|C_9| = 1$. In the appendix figures, we've shown corresponding centroids compared with the group average (the S&P 500 average) using the means of the first string letter of the moving window SAX representation; a visual representation of the SAX approximation. C_9 is Newmont Mining Corporation, a gold holding company. It seems nearly negatively correlated with the S&P 498, which isn't surprising. In bear markets, investors seek the security of gold, and opposite in bull markets. We also found that Newmont Mining Corporation was the farthest nearest neighbor. A clear indicator of its anomalous behavior. Company C_4 and C_5 are composed of medicinal supply companies, which seem to have done well through the holiday season, unlike others. The rest of the companies differ in their activity around the holiday season, and more importantly a period of financial turmoil driven by fears of recession. Clusters $3, 6, 1$ and 7 are more bullish than sectors $2, 10$ and 8.

Since our data does have an existing classification, the Global Industry Classification Standard (GICS), it is useful to compare how are clustering compares. Here we present a modified version of oft defined clustering features, centroid x_0, radius R, and diameter D of a cluster $X = \{x_1 \dots x_n\}$ [8].

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

$$R = \frac{\sum_{i=1}^{n} mindist(x_i, x_0)}{n} \tag{5}$$

$$D = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} mindist(x_i, x_j)}{\frac{n(n-1)}{2}} \tag{6}$$

R and D reflect the tightness of the cluster. We also want to capture the distance between cluster centroids. The results are shown in Table 1 of the Appendix. On average, the hierarchical clustering seems to outperform the GICS on these measures. Given these results, it's sensible to ask the extent to which each cluster is centered on one product class. 10 cluster results are shown in Table 2 of the Appendix. Most clusters seem fairly evenly spread out. The highest count C_1 does better than the GICS average, a surprising result given its high count. As far as product overlap, we get fairly mixed results. C_4 and C_5, which break up higher along the dendrogram, are highly related to the Health Care sector. When we produce 50 clusters, we continue to find most of our stocks in mixed sector clusters. Refer to Table 3 of the Appendix for results. In fact, in only a handful of cases, where the cluster size is larger than one, do we get 100% overlap and in all of those cases, there are at most two stocks in the cluster. The more common scenario is one where similarly priced stocks come from a number of industries, some more represented than others.

V. Conclusions and Future Research

We saw that with an equal number of clusters, our S&P 498 clustering produced clusters with better average price movement similarity than the GICS classification. The combination of this fact and the highly distributed nature of our stocks yields our conclusion, that industry sectors fail to adequately explain price movement. We think further research could be conducted along the lines of comparing our sectors with existing combination financial instruments such as mutual funds categorized by their growth, value, and risk.

References

[1] Z. Bodie, A. Kane, and A. J. Marcus, *Investments*, 7th ed. McGraw-Hill, 2007.

[2] B. W. Wah and M. Qian, "Constrained formulations and algorithms for stock-price predictions using recurrent fir neural networks," in *Eighteenth national conference on Artificial intelligence*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 2002, pp. 211–216.

[3] N. Da Costa, J. Cunha, and S. Da Silva, "Stock selection based on cluster analysis," *Economics Bulletin*, vol. 13, pp. 1–10, 2005.

[4] F. Rosen, "Correlation based clustering of the Stockholm Stock Exchange," Stockholm University, Tech. Rep., 2006.

[5] R. J. Yan and C. X. Ling, "Machine learning for stock selection," in *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2007, pp. 1038–1042.

[6] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. Association for Computing Machinery, 2003.

[7] L. W. Jessica Lin, Eamonn Keogh and S. Lonardi, "Experiencing sax: a novel symbolic representation of times series," *Data Mining and Knowledge Discovery*, vol. 15, Number 2/October, 2007, pp. 107–144, 2007.

[8] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann, 2005.