

IDENTIFYING TRENDS AND OUTLIERS IN NYSE INDUSTRIAL CATEGORIES

Andreea Dragut¹, Maria Dragut²

¹*Lab. Informatique Fondamentale, Univ. Aix-Marseille II, Ave. G. Berger, Aix-en -Pce, 13090 France*
E-mail: dragut@univmed.fr

²*Universitatea Tehnica de Constructii, Bd LACUL TEI 124, Bucuresti, 020396, Romania*

Abstract: The market tendency to move as a whole is accentuated in financial crisis times. Also normal consequence is the increased difficulty in attracting and retaining investors. Considering a stream of historical price data per each traded stock, we propose a linear time clustering algorithm which investigates to what extent stocks in a certain industry move as a group. To remove the overall market trend in we use the NYSE tick as a way of assessing whether buyers or sellers are more motivated. The algorithm thus build is monotonic and dissimilarities within industry groups is observed.

Keywords: linear on-line hierarchical classification, Lance-Williams algorithms

1. Introduction

In general clustering provides a synopsis of the data under scrutiny. It is used to obtain patterns or just to summarize information via cluster representatives. Most data mining algorithms do not work well for stream data due to their unique characteristics: the high dimensionality, high feature correlation, large amount of noise. Data mining of stream financial data has proven to be very effective and very profitable and stock clustering has applications to quantizing the effect of trends within and between industries, identifying misclassified stocks, and portfolio evaluation.

Every stock sold on the New York Stock Exchange is classified into an industrial category, or just called an “industry”. Stock market data tends to move as a whole, but they move with subtler trends on groups.

An attempt to cluster financial data in order to obtain a hierarchy of classes without any assumptions on the underlying structure of the data is the one of (Basalto, N. Bellotti R. De Carlo F. Facchi P. and Pascazio S. 2005) using a chaotic map clustering algorithm. They apply the algorithm to cluster thirty companies of the Dow Jones (DJ) market index , from 1998 to 2002. The correlation coefficients are computed for the logarithmic daily closure price variation. They show that categories as Conglomerates or Capital Goods are not behaving in an unitary way and are split among clusters.

But the historical record of prices shows very often that the movements of the stock within an industry are very not similar especially on crisis time. A logical explanation might be that during a crisis investors are more penalty than reward oriented. The industrial category of a company is obtained by identifying the area from which it derives the largest share of its revenue. A penalty oriented point of view would be the one of the increased difficulty in attracting and retaining investors as emphasized by the 2008 survey done for the NYSE EURONEXT 2009 CEO Report: Managing During Economic Turbulence.

In this paper we consider a stream of historical price data per each traded stock and we propose a linear time clustering algorithm which investigates to what extent stocks in a certain industry move as a group. The measure between stocks is constructed using the NYSE tick indicator which is the net number of stocks rising or declining in price by subtracting from the upticks all the downticks at a given point during the day in NYSE. Tick statistics are available also for Nasdaq and AMEX.

2. The data and tick measure

Let $s_i(t)$ denote the price of stock i at time t . During trading hours, the price of a stock is constantly fluctuating. The price of a stock does not necessarily reflect the revenue or size of a company. As in (Gavrilov,

D. Anguelov, P. Indyk, and R. Motwani, 2000), we consider the percentage change $P_i(t)$ in a stock i as good comparative measure of stock performance at a fixed time. $P_i(t) = 100 \frac{s_i(t+1) - s_i(t)}{s_i(t)}$

To remove the overall market tendency such that subtler trends can be studied the data is normalized. There are two basic approaches to normalizing stock market data: stock-based normalization and time-based normalization. In stock-based normalization, one calculates the mean percent change as well as standard deviation for each stock. Then one normalizes the percentage change $P_i(t)$. In time-based normalization of $P_i(t)$, one uses the mean percent change for a time interval across all stocks and the standard deviation for that time period.

Another popular pre-processing step is to attempt to reduce the dimensionality of the data by subsequence clustering using the sliding windows technique. We are not choosing this approach and we prefer to propose a linear time algorithm for clustering the complete sequences, since in (Keogh E, Lin J, Truppel W 2003) this approach did come under scrutiny. The research of (Chen J.R. 2007) confirmed that the subsequence clustering together with the use of Euclidean distance and of a cluster centroid did not lead to a sensible clustering outcome. To compensate, (Chen J.R. 2007) introduced a new mixed measure.

On the stock market, buyers want to buy at low prices and sellers to sell at high prices. The bid price is the highest price a buyer wants to pay, and the offer price the lowest a sellers accepts. Offer minus bid is called spread: the more liquid the stock, the narrower its spread. The book contains buy-orders below the bid and sell-orders above the offer -- for more patient players -- waiting for price moves. The less patient ones place market orders: buyers accept the offer, typically trading on upticks; conversely, sellers accept the bid, thus on downticks. The NYSE Tick $N(t)$ is the number of upticking stocks minus the downticking ones: calculated by NYSE every six seconds, it can be used as a short-term trend indicator.

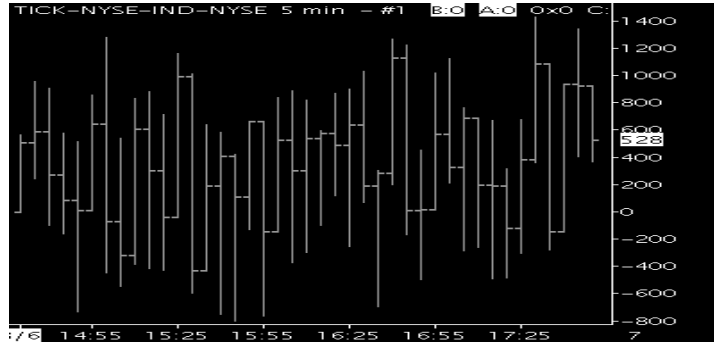


Fig. 1. 5-min sample of NYSE Tick statistic

Another option for removing the overall market influence is to normalize the data using the NYSE tick as a long-time indicator. The cumulative as well as the net NYSE tick are both biased.

3. Results

In our paper, we adapt to our problem a fast on-line hierarchical clustering algorithm from (Dragut Nichitiu 2003), taking $O(n)$ time for two classes, and $O(n \cdot k)$ for k classes.

We prove for this algorithm the *monotonicity* (also formally defined later): the dissimilarity between classes increases with each merge. This means the classification is improving the separation of classes on the way up in the hierarchy. This property is very useful for algorithms, because even if they do not have the ISED property, one could rerun them in some classes to better split them (if the algorithm is fast); at least one is sure the already separated classes have been separated quite well. Given the fact that no known algorithms have the ISED property, if one has to choose between a numerical clustering algorithm for which it has been proven that it does not have the monotonicity property, and another one, for which it has been proven that it does have the monotonicity property, the second is better.

Definition An $LW(\alpha, \beta, \gamma, p)$ algorithm Lance-Williams characterized by the inter-class distance $\tilde{d} : N \times N \rightarrow \mathbf{R}$ a dissimilarity index on N is a function is called monotonic if for any three classes A_k, A_i, A_j the following inequality takes place:

$$\tilde{d}(A_i \cup A_j, A_k) \geq \tilde{d}(A_i, A_j).$$

Suppose the completed time series to be classified, that is the elements of N a NYSE industrial category, arrive in a certain order, and we have to classify them as soon as they are observed. Let us also suppose that each element can be identified with a vector of \mathbf{R}^s . Let $(X_1(t), \dots, X_n(t))$ be the observed sequence at the moment t , and each $X_i(t) = (X_{i1}(t), \dots, X_{is}(t))$.

Since $N \subset \mathbf{R}^s$, a convenient representation for a class of q elements of N is the vector which represents the gravity center of the q elements.

Definition (Everitt, B. 1993) *The statistical distance among two classes A_i , A_j is defined as*

$$d_a(A_i, A_j) = \Delta E_{A_i, A_j}^2 = e_{A_i \cup A_j}^2 - e_{A_i}^2 - e_{A_j}^2,$$

where $e_A^2 = \sum_{i=1}^{|A|} \sum_{j=1}^s (x_{ij} - m_j^A)^2$, and $m_j^A = \frac{1}{|A|} \sum_{i=1}^{|A|} X_{ij}$.

Remark *The reasons for choosing this special distance lie with its special properties. It so happens that the characterization of a class only through the position of its centroid does not offer sufficient information about the positions of the class elements: to a same centroid, very close elements or very remote ones in \mathbf{R}^s may equally correspond*

Happily, we can also compute this distance with a low computational effort, due to Everitt. For two classes A , A_j , we have

$$d_a(A_i, A_j) = \frac{|A| \cdot |A_j|}{|A| + |A_j|} \frac{\sum_{k=1}^s (m_k^A - m_k^{A_j})^2}{\sum_{k=1}^s (m_k^A - m_k^{A_j})^2} = \frac{|A| \cdot |A_j|}{|A| + |A_j|} d(X_A, X_{A_j})$$

where $X_A = \frac{1}{|A|} \sum_{X \in A} X$.

Proposition *The on-line hierarchical algorithm using as distance between two classes the statistical distance divided by the NYSE tick at the moment of a change in a time series is monotonic.*

Proof We use the induction on the number k of the objects to be classified.

Let $k = 3$, and let $X_1(t), X_2(t), X_3(t)$ the objects to be classified. From the step 1 of the algorithm we have three possibilities to classify them at the time instant $t \geq 1$:

1. $\{X_1(t), X_2(t)\}, \{X_3(t)\}$ if $d_a(\{X_1(t)\}, \{X_2(t)\}) \leq \min\{d_a(\{X_1(t)\}, \{X_3(t)\}), d_a(\{X_2(t)\}, \{X_3(t)\})\}$
2. $\{X_1(t), X_3(t)\}, \{X_2(t)\}$ if $d_a(\{X_1(t)\}, \{X_3(t)\}) < d_a(\{X_1(t)\}, \{X_2(t)\})$,
 $d_a(\{X_1(t)\}, \{X_3(t)\}) \leq d_a(\{X_2(t)\}, \{X_3(t)\})$ and
3. $\{X_2(t), X_3(t)\}, \{X_1(t)\}$ if $d_a(\{X_2(t)\}, \{X_3(t)\}) < d_a(\{X_1(t)\}, \{X_2(t)\})$, $d_a(\{X_2(t)\}, \{X_3(t)\}) \leq d_a(\{X_1(t)\}, \{X_3(t)\})$

Since an elementary computation from (Everitt, 1993) gives that

$$d_a(A_k, A_i \cup A_j) = \frac{|A_i| + |A_k|}{|A_i| + |A_j| + |A_k|} d_a(A_k, A_i) + \frac{|A_j| + |A_k|}{|A_i| + |A_j| + |A_k|} d_a(A_k, A_j) - \frac{|A_k|}{|A_i| + |A_j| + |A_k|} d_a(A_i, A_j)$$

we have

$$\begin{aligned}
 d_a(\{X_1(t), X_2(t)\}, \{X_3(t)\}) &= \frac{2}{3} d_a(\{X_1(t)\}, \{X_3(t)\}) + \frac{2}{3} d_a(\{X_2(t)\}, \{X_3(t)\}) \\
 &\quad - \frac{1}{3} d_a(\{X_1(t)\}, \{X_2(t)\}) \\
 &\geq d_a(\{X_1(t)\}, \{X_2(t)\}), \\
 d_a(\{X_1(t), X_3(t)\}, \{X_2(t)\}) &= \frac{2}{3} d_a(\{X_1(t)\}, \{X_2(t)\}) \\
 &\quad + \frac{2}{3} d_a(\{X_2(t)\}, \{X_3(t)\}) - \frac{1}{3} d_a(\{X_1(t)\}, \{X_3(t)\}) \\
 &\geq d_a(\{X_1(t)\}, \{X_3(t)\})
 \end{aligned}$$

The third case can be treated similarly. Assume that the property holds for k objects, and we want to prove it for $k+1$ objects. Let A_1 and A_2 the two classes situated at the first level of the hierarchical tree constructed by the algorithm for the first k objects. From the step i we have:

1. $(A_i(t) \cup A_j(t)), \{X_{k+1}(t)\}$ if $d_a(A_i(t), A_j(t)) \leq \min(d_a(A_j(t), \{X_{k+1}(t)\}), d_a(A_i(t), \{X_{k+1}(t)\}))$,
2. $(A_j(t) \cup \{X_{k+1}(t)\}), A_i(t)$ if $d_a(A_j(t), \{X_{k+1}(t)\}) < d_a(A_i(t), A_j(t)), d_a(A_j(t), \{X_{k+1}(t)\}) \leq d_a(\{X_{k+1}(t)\}, A_i(t))$
3. $(A_i(t) \cup \{X_{k+1}(t)\}), (A_j(t))$ if $d_a(A_i(t), \{X_{k+1}(t)\}) < d_a(A_i(t), A_j(t)), d_a(A_i(t), \{X_{k+1}(t)\}) \leq d_a(\{X_{k+1}(t)\}, A_j(t))$ Since

$$\begin{aligned}
 d_a(A_i(t) \cup A_j(t), \{X_{k+1}(t)\}) &= \frac{|A_i(t)| + 1}{|A_i(t)| + |A_j(t)| + 1} d_a(A_i(t), \{X_{k+1}(t)\}) \\
 &\quad + \frac{|A_j(t)| + 1}{|A_i(t)| + |A_j(t)| + 1} d_a(A_j(t), \{X_{k+1}(t)\}) \text{ we have} \\
 &\quad - \frac{1}{|A_i(t)| + |A_j(t)| + 1} d_a(A_i(t), A_j(t))
 \end{aligned}$$

$d_a(A_i \cup A_j, \{X_{k+1}(t)\}) \geq d_a(A_i, A_j)$. The other two cases can be treated similarly.

5. Conclusion

Several observed differences with respect to the industrial category are that since 2007 Time Warner Inc. Consumer Discretionary /Media behave more as Information Technology stocks, Eastman Kodak - Consumer Discretionary /Leisure Equipment and Products General Electrics – Industrial Conglomerates/Industrials, General Motors - Consumer Discretionary/Automobiles are behaving like Financials. Confirmed by their recent meltdown, the clusters showed the market was aware of their risk exposure and clustered them by penalizing a shared risk with other industries and by rewarding the primary source of revenues.

On the contrary in the Health Care industry the core companies as well as the Pharmaceuticals, Biotechnology and Life Science behaved homogenously.

References

- Basalto, N.; Bellotti, R.; De Carlo, F.; Facchi, P. and Pascazio, S. 2005. Clustering stock market companies via chaotic map synchronization, *Physica A: Statistical Mechanics and its Applications* 345 (Issues 1–2, 1), 196–206.
- Chen, J. R. 2007. Making clustering in delay-vector space meaningful, *Knowledge and Information Systems* 11(Number 4, April): 369–385.
- Dragut, A and Nichitiu, C. 2004. A monotonic on-line linear algorithm for HAC, *Information and Technology Management* 5: 111–141.
- Everitt, B. 1993. *Cluster Analysis*. John Wiley & Sons, Inc.
- Gavrilov, M.; Anguelov, D.; Indyk, P. and Motwani, R. 2000. Mining the Stock Market: Which Measure is Best? *Proc. of the KDD*, 487–496.
- Keogh, E.; Lin J., Truppel, W. 2003. Clustering of time series subsequences is meaningless: implications for previous and future research, in *Procs of the international conference of data mining*, 19–22.