Modified k-means algorithm for clustering stock market companies

Article · January 2007

2 authors:

Parviz Rashidi
Umutara Polytechnic
2 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Morteza Analoui
Iran University of Science and Technology
123 PUBLICATIONS **363** CITATIONS
SEE PROFILE

IDMC'07
20-21 Nov.2007
Amir Kabir University

Modified k-means algorithm for clustering stock market companies

Parviz Rashidi[1] p_rashidi@comp.iust.ac.ir

M.Analoui[1] Analoui@iust.ac.ir

Javad Azizmi[1] ja_azimi@comp.iust.ac.ir

[1]Iran University of Science and Technology, Computer Engineering Department

Abstract

In recent years, there has been a lot of interest in the database community in mining time series data, especially in finance markets. Partitioning assets into natural groups or identifying assets with similar properties are natural problems in finance. In this paper, we proposed a *modified k-means clustering algorithm* to cluster stock market companies, based on similarity measure between time series. This algorithm utilize *maximum information compression (MIC) index* as similarity measure for clustering them and its comparison with two other similarity measures, namely *correlation coefficient* and *least-square regression error* are made. Appling this algorithm leads to a natural partition of the data, as companies belonging to the same industrial branch are often grouped together. This algorithm is applied to the analysis of the Dow Jones (DJ) index companies, in order to identify similar temporal behavior of the traded stock prices. The identification of clusters of companies of a given stock market index can be exploited in the portfolio optimization strategies.

Keywords: clustering, feature similarity, feature selection, stock market index.

1. Introduction

There has been a lot of interest in the database community in mining time series data, especially in finance markets. Partitioning assets into natural groups or identifying assets with similar properties are natural problems in finance [1, 3 and 4], so the objective of this attention is to understand the underlying dynamics which rules the company's stock prices. In particular, it would be useful to find, inside a given stock market index, groups of companies, which have similar temporal behavior. This behavior can be extracted from the price of company at last periods and apply it to cluster similar companies with each other. To do clustering, hierarchical and partitional clustering algorithms [5] can be used and Comparison between hierarchical single-linkage algorithm and partitional k-means algorithm are presented in [1], which use Euclid distance as similarity measure. A. S. Weigend [4] described the strengths and weakness of each of the method on financial data. N.Basalto and other [3] clustered stock market companies via chaotic map synchronization algorithm. In this paper, we used idea of feature clustering [2] and proposed a *modified k-means clustering algorithm*, which use similarity measure between historical data of stock market companies to cluster them. In the standard k-means algorithm, Euclid distance [5] is used for clustering pattern in input space and its application on financial data are presented in [1] , here we used *maximum information compression index (MIC)* [2] as similarity measure and its comparison with two other similarity measures, namely *correlation coefficient* [6] and *least-square regression error* [7] is made. Presented result shows that this proposed algorithm, which use *MIC* index as similarity measure, are more suitable than standard k-means in finding natural partition of data, as the companies belonging the to the same industrial branch are often grouped together. Because of Sensitivity to initial seed points in k-means, we selected initial seed point with k-nn algorithm presented in [2].

P. Mitra, C.A. Murthy and K. Pal [2] used k-nn algorithm for feature clustering based on those similarity measures and then utilize it for feature selection, but presented k-nn are not suitable for our work. We modified other well-known clustering algorithm namely k-means to cluster companies. In this work, we mapped historical data of stock market companies to feature clustering problem, so we are able to use any feature clustering method for clustering stock market companies and similar markets. In this mapping each day of historical data in last period are considered as features and selected companies are our pattern.(Ex: if we have 1000 record of historical data of 30 companies, we have 30 patterns which have 1000 feature). Main objective of this work is identifying similar temporal behavior of the traded companies in stock market and grouping them. Presented algorithm is applied to cluster selected companies of Dow Jones (DJ) index companies. The identification of clusters of companies of a given stock market index can be exploited in the portfolio optimization strategies.

This paper is organized as follows: in Section 2 we give a brief review of the similarity measures. Section 3 deals with our work and analysis of the companies' stock prices to cluster them. Finally, some conclusions are drawn in Section 4.

2. Similarity measure

In this section we discus similarity measure between companies, based on linear dependency between them. There are broadly two possible approaches for measuring similarity between variables. One is to non-parametrically test the closeness of the probability distribution between variables [8], however this approach are not suitable for our work. Another approach is to measure the amount of functional (linear or higher) dependency between companies. There are several benefits to choosing linear dependency as a similarity measure instead of higher dependency [2]. In below we discuss some linear similarity measures.

- *Maximal information compression index(λ)* [2] : this index defined as

$$\lambda(x,y) = (\mathrm{var}(x) + \mathrm{var}(y) - \sqrt{(\mathrm{var}(x) + \mathrm{var}(y))^2 - 4\,\mathrm{var}(x)\,\mathrm{var}(y)(1 - \rho(x,y)^2}\,)/2$$

The value of λ is zero when two features are linearly dependent and increases as amount of dependency decrease. The significance of λ can also be explained geometrically in term of linear regression. It can be easily shown[8] that value of λ is equal to the sum of the square of the perpendicular distance of the points(x ,y) to the best fit line $y = \hat{a} + \hat{b}x$, obtained by minimizing the some of the square perpendicular distance. The coefficient of such a best fit line are given by $\hat{a} = \bar{x}\cot(t) + \bar{y}$ and $\hat{b} = -\cot(t)$, where $t = 2\tan^{-1}(\dfrac{2\,\mathrm{cov}(x,y)}{\mathrm{var}(x)^2 - \mathrm{var}(y)^2})$.

- *Correlation coefficient (ρ) [6]*: one of the well-known measures of similarity between two random variables is the Correlation coefficient ρ. Correlation coefficient ρ between two random variable x and y is defined as. $\rho(x,y) = \dfrac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$

- *Least square regression Error (e) [7]:* another measure of the degree of linear dependency between two variable x and y is the error in predicting y from linear model $y = a + bx$, a and b are regression coefficient obtained by minimizing the mean square error. $e(x,y)^2 = \dfrac{1}{n}\sum(e(x,y)_i)^2$, $e(x,y)_i = y_i - a - bx_i$, The coefficient are given by $a = \bar{y}$ and $b = \dfrac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$.

The measure λ possesses several desirable properties like symmetry, sensitivity to scaling and invariance to rotation, but the measure e is not symmetric and ρ is invariant to scaling. Experiments result shows that the λ measure are suitable than other measures.

3. Our Work

In this section we describe our work, clustering stock market companies. To this purpose, we use idea of feature clustering [2] and modified standard k–means [5] clustering algorithm based on one of the feature similarity measure presented in section 2. In feature clustering method each pattern have a few features, so in this work selected companies are our patterns and daily closure price are features of them. Presented similarity measures between features are computed for daily price variation time series. The daily price variation is given by $y_i(t) = P_i(t+1) - P_i(t)$, Where P_i (t) is the closure price of stock *i* at time *t*. it is shown [1] that first derivative was much more informative than the original series.

P.Mitra and other [2] used k-nearest neighboring algorithm for clustering feature, but that algorithm are not suitable for us, so we modified k-means clustering algorithm. Modification of this algorithm is based on calculating similarity measure between patterns, in standard version of k-means, Euclid distance are used as similarity between patterns, but here we used presented similarity measure in section 2 as a distance between companies. For solving the sensitivity to initial seed point's problem of k-means algorithm, we used k-NN [2], which select patterns as seed point that has a compact subset.

The algorithm can be stated as follow:

- Begin with k cluster, each consisting of one of samples selected by k-nn. (In standard k-means first k samples considered as centroids).
- For each remaining n-k samples, find the centroid nearest it based on similarity measure presented in section 2, put the sample in the cluster identified with this nearest centroid. After each sample is assigned, recompute the centroid of the altered cluster.
- Go through the data a second time. For each sample, find the centroid nearest it. Put the sample in the cluster identified with this nearest centroid. (During this step, don't recompute any centroids).

We applied this algorithm to cluster the companies of the Dow Jones (DJ) Stock market index. Including N=30 stocks [9], whose names are listed in Table1.

The similarity measure is computed for daily closure price variation on time series from 1998 to 2002. The daily closure price variation is given by $y_i(t) = P_i(t+1) - P_i(t)$, Where P_i(t) is the closure price of stock *i* at time *t*. The result of applying standard k-means algorithm are presented in Table2 and result of applying the proposed algorithm are shown in Table3.

Table1: Dow Jones stock market companies

Company Name	***Abbr.***	***Group***	***Company Name***	***Abbr.***	***Group***
Alcoa Inc	AA	BM*	International Paper	IP	BM
American Express Co.	AXP	Fin	Johnson & Johnson	JNJ	H
Boeing	BA	CG	JP Morgan Chase	JPM	Fin
Citigroup	C	Fin	Coca Cola Inc	KO	CNC
Caterpillar	CAT	CG	McDonalds Corp	MCD	S
DuPont	DD	BM	Minnesota Mining	MMM	Cong
Walt Disney	DIS	S	Philip Morris	MO	CNC
Eastman Kodak	EK	CC	Merck & Co	MRK	H
General Electrics	GE	Cong	Microsoft	MSFT	T
General Motors	GM	CC	Procter & Gamble	PG	CNC
Home Depot	HD	S	SBC Communications	SBC	S
Honeywell International	HON	CG	AT&T Gamble	T	S
Hewlett-Packard	HPQ	T	United Technology	UTX	Cong
International Business Machine	IBM	T	Wal-Mart Stores	WMT	S
Intel Corporation	INTC	T	Exxon Mobil	XOM	E

*Basic Materials (BM), Financial (Fin), Capital Goods (CG), Services(S), Consumer Cyclical (CC), Conglomerates (Cong), Technology (T), Healthcare (H), Consumer Non-Cyclical (CNC),Energy(E)

Table2: result of standard k-means algorithm

Cluster1	**Cluster2**
• CAT (CG) • IP (BM) • MCD, SBC(S) • MO (CNC)	Other companies (83%)

Table3: result of the proposed modified k-means algorithm

Similarity Measure / **Clusters**	***Maximal information compression index(λ) [2]***	***Least square regression Error(e)***	***Correlation coefficient (ρ)***
Cluster 1	• AA , DD, IP (BM) • BA, CAT, HON (CG) • AXP,C, JPM (Fin) • DIS , HD (S) • EK, GM (CC) • GE (Cong) • HPQ, IBM, INTC (T) • JNJ (H)	• AA, DD (BM) • MMM,UTX(Cong) • MO, PG (CNC) • MRK (H) • MSFT (T) • SBC,T,WMT(S) • XOM (E)	• AA, DD (BM) • AXP, C (Fin) • BA, CAT (CG)
Cluster 2	• KO, MO, PG(CNC) • MCD, SBC, T ,WMT (S) • MMM, UTX (Cong) • MRK (H) • MSFT (T) • XOM (E)	• AXP,C, JPM (Fin) • BA,CAT, HON (CG) • DIS, HD, MCD (S) • EK, GM (CC) • GE (Cong) • HPQ, IBM,INTC (T) • IP (BM) • JNJ (H) • KO (CNC)	Other companies (80%)

The presented result shows that the proposed algorithm, which use λ, *e and* ρ as similarity measure are suitable than the standard k-means in clustering stock market time series, because Applying the standard k-means algorithm with similarity measure D_e clustered companies in two clusters like other measure but most of the companies (83%) are grouped in same cluster and companies in other cluster aren't related to each other and this case is same as the state all company are grouped in one cluster and have a drawback of chaining effect . It can be concluded from the presented result that similarity measure λ are better than other measure because with this measure most of the companies belonging to the same industrial branch are often clustered in same group. It's shown that based of this measure all company in industrial branches Basic Material (BM), Financial (Fin), Capital Goods (CG), Consumer Cyclical (CC) and Consumer Non-Cyclical (CNC), and most of the companies in Services(S) and Technology (T) are grouped with each other.

4. Conclusion

In this work we used idea of feature clustering and proposed *modified k-means algorithm* to cluster stock market companies, which utilize presented similarity measure in section 2 as distance between companies. K-means sensitivity to initial seed points are solved by selecting seed points with k-nn. The presented result shows that the proposed algorithm, which utilize *Maximal information compression index (λ)* as similarity measure is more suitable than standard one to find natural partition of the data, as the companies belonging to the same industrial branch are often grouped together. The clustering output can be exploited in portfolio optimization strategies.

Reference:

[1]. M. Gavrilov, D. Anguelov, P. Indyk and R. Motwani."Mining the Stock Market: Which Measure is Best?", Proc. Sixth Int. Conf. Knowledge Discovery and Data Mining (KDD), 487-496, 2000.

[2]. P. Mitra, C.A. Murthy and K. Pal , "unsupervised feature clustering using feature similarity", IEEE Transaction on pattern analysis and machine intelligence ,vol.24 , No 3, 2002.

[3]. N. Basalto, R. Bellotti, F. De Carlo, P. Facchi and S. Pascazio, "Clustering stock market companies via chaotic map synchronization", Elsevier ,Physica A 345 196–206, 2005.

[4]. A.S. Weigend, "Data Mining in Finance: Report from the Post-NNCM-96 Workshop on Teaching Computer Intensive Methods for Financial
Modeling and Data Analysis", Proc. Fourth International Conference on Neural Networks in the Capital Markets NNCM-96, p. 399-411, 1997.

[5]. E.Gose and R.jhonsonbaugh, "Pattern recognition and image analysis", prentice hall Inc, 1996 .

[6]. C.R. Roa, "linear statistical inference and its application", John Wiley, 1973.

[7]. R.L. Plackett, "The discovery of the method of least squares.Biometrika", 59, 239–251, 1972.

[8]. P.A. Devijver and J. Kittler, "Pattern recognition: A Statistical approach". Englewood Cliffs: prentice Hall, 1982.

[9]. "Yahoo! Financial." http://finance.yahoo.com.