時系列パターン抽出に基づく株売買ルール生成手法の評価

Evaluation of a Stock Trading Rule Generation Method based on Temporal Pattern Extraction

阿部秀尚 [1]　杉山喜昭 [2]　山口高平 [2]

Hidenao Abe[1], Yoshiaki Sugiyama[2], and Takahira Yamaguchi[2]

[1] 島根大学

[1]Shimane University

[2] 慶應義塾大学

[2] Keio University

Abstract: In this paper, we present an evaluation of a temporal rule generation method for trading dataset from the Japanese stock market. Temporal data mining is one of key issues to get useful knowledge from databases. To get more valuable rules for users from a temporal data mining process, we have developed a rule generation method which consists of temporal pattern extraction methods and rule induction algorithms. Using this method, we have done a case study to evaluate temporal rules from a Japanese stock market database for trading. Based on the result, we discuss about a way to utilize our rule generation method more effectively.

1. はじめに

近年，データベースシステムなどに蓄積されたデータを体系的に分析し，知識を得るデータマイニングが広く知られるようになってきた．特に株価や商品市場のように時系列でデータが得られる分野では，時系列の類似性や規則性を発見する手法[1][2]や時系列ルールマイニング[3]が提案されてきた．一方，専門知識を持たない個人投資家は，株価や多くのテクニカル指標から売買に有用な判断材料を得るために多大なコストが必要となり，高度な知識や経験の習得を支援する手法が求められている．

本研究では，時系列の株価とテクニカル指標から成るデータから，個人投資家の売買判断を支援するルールの生成を行うツールの開発を目的としている．本ツールでは，時系列データからクラスタリングによる代表パターンの抽出に基づき，決定木学習やルール学習によって生成を行う．出力されるルールは，売買判断に必要となる株価やテクニカル指標の組み合わせとなり，これらは時系列と併せて利用者に提示される．このため，テクニカル分析やチャート分析の専門家には新たな視点からの売買判断規則を提示することを可能とし，そうでない利用者にも有用な規則を提供することとなる．

本稿では，生成されたルールの予測過程における時系列パターンを同定する精度，およびルール集合全体での売買判断の予測精度を評価する．さらに，売買判断をより有効に行うための要因について検討を行う．

2. 時系列ルールマイニングツール

本研究で開発を行う時系列ルールマイニングツールの概観を図 1 に示す．

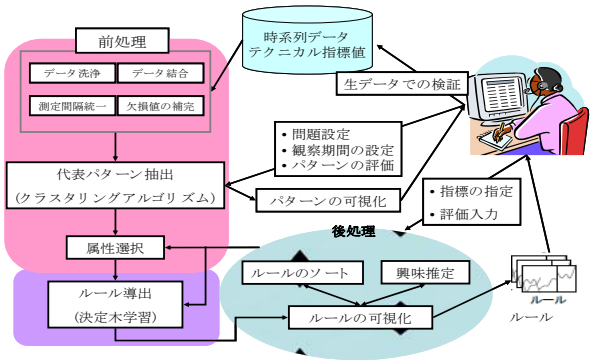

図 1　システム概要

株価データについての時系列ルールの作成手順は，前処理において，テクニカル指標値の時系列データからサブシーケンスを切り出し，クラスタリングアルゴリズムを用いることによって代表的なパターンから成る訓練データを作成する．この訓練データに

売買判断となるクラスを加え，決定木学習やルール学習手法を適用して売買判断ルールを出力する．後処理においては，時系列ルールを可視化して利用者に示し，必要に応じて生データの確認が行えるインタフェースを提供する．

2.1 前処理：代表パターン抽出に基づく訓練データの作成

前処理では，最初にテクニカル指標値の時系列データからウィンドウサイズに応じたサブシーケンスを切り出す．ウィンドウサイズはサブシーケンスを切り出すための期間であり，観察期間を何日もうけるか，を意味する．次に，サブシーケンスから成るデータから各テクニカル指標のデータを取り出し，クラスタリングを適用，任意の数の代表パターンを抽出する．以上の手順の概観を図 3 に示す．

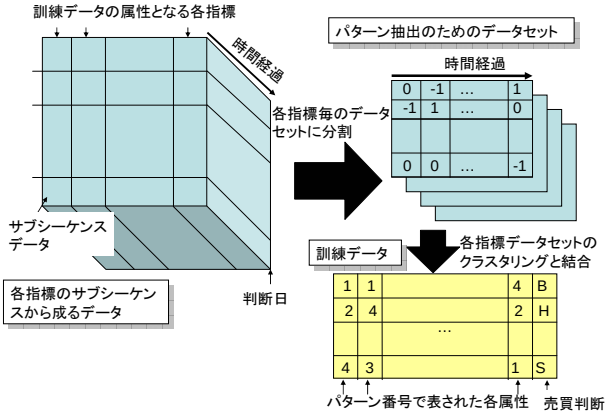

図２　訓練データ作成手順の概観

訓練データは，テクニカル指標ごとに抽出された代表パターンを属性とし,期間末日の株式売買における正しい判断をクラスとする．ここで,正しい判断とは設定された百分率(x)に対して，その日から n 日間で終値が x%増加したら「買い」, x%減少したら「売り」，そのいずれでもなかったら「保持」を選択することをいう．ただし，「買い」と「売り」を出すための条件とをともに満たした場合は，増減率の大きいほうを採用する．

3. 時系列ルールの精度評価

本節では，株ロボ[4]から取得した 2006 年 1 月 5 日～2006 年 5 月 31 日までのデータを用いて，時系列ルール作成と作成されたルールの予測精度の評価について述べる．

本実験で対象とした銘柄は，銀行・金融業から 5 銘柄（三菱 UFJ フィナンシャルグループ，みずほフィナンシャルグループ，三井住友フィナンシャルグループ，クレディ・セゾン，オリックス）と通信業 4 銘柄（NTT，NTT ドコモ，KDDI，ソフトバンク）である．また，正しい売買判断として 20 日後に 5%の上昇を「売り」，下降を「買い」，それ以外を「保持」と設定した. 各銘柄のクラス分布を表 1 に示す．

表１　各銘柄のクラス分布

銀行･金融業	買い	売り	保持	通信業	買い	売り	保持
セゾン	37	53	9	NTT	27	32	40
オリックス	43	40	16	KDDI	42	39	18
三菱UFJFG	0	50	49	NTTドコモ	19	29	51
三井住友FG	6	27	66	ソフトバンク	23	69	7
みずほFG	38	31	30				

本実験での観察期間は，75 日として，クラスタリング手法として K-Means と EM アルゴリズムによるクラスタリングを用いた．

3.1 各属性の詳細

本実験で利用したテクニカル指標とそれらの計算式を表 2 に示す．基本的なパラメータとして，長期の移動平均は 26 日, 短期の移動平均は 9 日を用いる．また，一目均衡表からの値については長期を 26 日，短期を 9 日として算出している．

株価データ（始値(opening), 高値(high), 安値(low), 終値(closing)）および出来高(Volume)，テクニカル指標である volumeRatio と RSI，Momentum は時系列データから得られた各日の値を用いた．それ以外の指標は買い(Buy)であれば 1，売り(Sell)であれば-1，それ以外は 0 を割り当てた．これらの値は，売買判断日 t について計算され，訓練データの各データとなる．

3.2 時系列パターンに基づくルールによる予測

時系列ルールによる予測では，通常のルール予測とは異なり，テストデータのパターンを同定する必要がある．この処理はシステム内で行えるため，利用者への分かりやすさよりもパターンの予測精度が重要となるため，メタ学習スキームやパターン識別アルゴリズムが適用可能である．訓練データの時系列パターンに基づくテストデータのパターン同定の処理について，概観を図 3 に示す．

表 2 訓練データにおける属性とその計算式

Attribute name		Description
RAW	opening	opening price of the day (O_t)
	high	Highest price of the day (H_t)
	low	Lowest price of the day (L_t)
	closing	Closing price of the day (C_t)
	Volume	Volume of the day (V_t)
TREND INDICES	Moving Average	Buy: if $SMA_t - LMA_t < 0 \cap SMA_{t-1} - LMA_{t-1} > 0$,Sell: if $SMA_t - LMA_t > 0 \cap SMA_{t-1} - LMA_{t-1} < 0$ Where $SMA_t = (C_t + C_{t-1} + \ldots + C_{t-12})/13$, and $LMA_t = (C_t + C_{t-1} + \ldots + C_{t-25})/26$
	Bolinger Band	Buy: if $C_t \geq (MA_t + 2\sigma) \times 0.05$, Sell: if $C_t \leq (MA_t - 2\sigma) \times 0.05$ where $MA_t = (C_t + C_{t-1} + \ldots + C_{t-24})/25$
	Envelope	Buy: if $C_t \geq MA_t + (MA_t \times 0.05)$, Sell: if $C_t \leq MA_t - (MA_t \times 0.05)$
	HLband	Buy: if $C_t < LowLine_{t-10\,days}$, Sell: if $C_t > HighLine_{t-10\,days}$
	MACD	Buy: if $MACD_t - AvgMACD_{t-9days} > 0 \cap MACD_{t-1} - AvgMACD_{(t-1)-9days} < 0$ Sell: if $MACD_t - AvgMACD_{t-9days} < 0 \cap MACD_{t-1} - AvgMACD_{(t-1)-9days} > 0$ Where $MACD_t = EMA_{t-12\,days} - EMA_{t-26\,days}$, $EMA_t = EMA_{t-1} + (2/range + 1)(C_{t-1} - EMA_{t-1})$
	DMI	Buy: if $PDI_t - MDI_t > 0 \cap PDI_{t-1} - MDI_{t-1} < 0$, Sell: if $PDI_t - MDI_t < 0 \cap PDI_{t-1} - MDI_{t-1} > 0$ Where $PDI = \sum_{i=t-n}^{t}(H_i - H_{i-1}) \div \sum_{i=t-n}^{t} TR_i \times 100$, $MDI = \sum_{i=t-n}^{t}(L_i - L_{i-1}) \div \sum_{i=t-n}^{t} TR_i \times 100$, $TR_t = \max\{(H_t - C_{t-1}), (C_{t-1} - L_t), (H_t - L_t)\}$
	volumeRatio	$VR_t = \{(\sum_{i=t-25, H_i > L_i}^{t} V_i + \sum_{i=t-25, H_i = L_i}^{t} V_i)/(\sum_{i=t-25, H_i < L_i}^{t} V_i + \sum_{i=t-25, H_i = L_i}^{t} V_i)\} \times 100$
	RSI	$RSI_t = 100 - 100/\{\sum_{i=t-13, C_{i-1} < C_i}^{t}(C_{i-1} - C_i)/\sum_{i=t-13, C_{i-1} > C_i}^{t}(C_{i-1} - C_i) + 1\}$
	Momentum	$M_t = C_t - C_{t-10}$
	Ichimoku1	Buy: if $C_{t-1} < RL_{t-9days} \cap C_t > RL_{t-9days}$, Sell: if $C_{t-1} > RL_{t-9days} \cap C_t < RL_{t-9days}$ Where $RL_{t-9days} = average\ (\max(H_i) + \min(L_i))\ \ (i = t-8, t-7, \ldots t)$
	Ichimoku2	Buy: if $C_{t-1} < RL_{t-26days} \cap C_t > RL_{t-26days}$, Sell: if $C_{t-1} > RL_{t-26days} \cap C_t < RL_{t-26days}$ Where $RL_{t-26days} = average\ (\max(H_i) + \min(L_i))\ \ (i = t-25, t-24, \ldots t)$
	Ichimoku3	Buy: if $RL_{(t-2)-26days} < RL_{(t-2)-9days} \cap RL_{(t-1)-26days} > RL_{(t-1)-9days} \cap RL_{(t-1)-26days} < RL_{t-26days}$ Sell: if $RL_{(t-2)-26days} > RL_{(t-2)-9days} \cap RL_{(t-1)-26days} < RL_{(t-1)-9days} \cap RL_{(t-1)-26days} > RL_{t-26days}$
	Ichimoku4	Buy: if $C_t > AS1_{t-26} \cap C_t > AS2_{t-26}$, Sell: if $C_t < AS1_{t-26} \cap C_t < AS2_{t-26}$ Where $AS1_t = median\ (RL_{t-9days} - RL_{t-26days})$, $AS2_t = (\max(H_t) - \min(L_t))/2\ \ (i = t-51, t-50, \ldots t)$

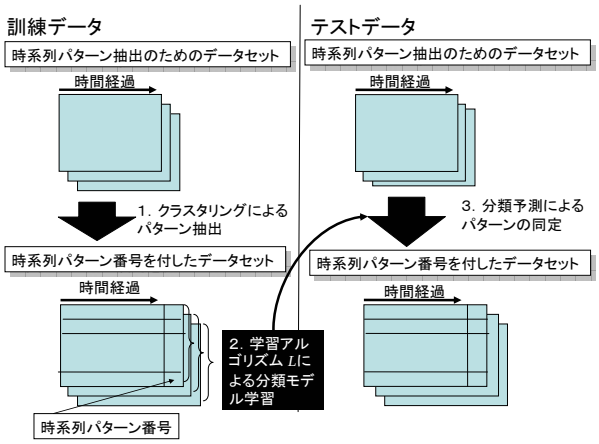

図 3 時系列パターンの同定

本実験では，パターンの同定手法として Boosted C4.5[5][6]（Weka[7]に実装された AdaBoost と J4.8）を利用した．５銘柄と２種類のクラスタリング手法について，同一銘柄でのパターンの予測精度を表 3 に示す.

表 3 同一銘柄での時系列パターン予測精度(%)

金融・銀行業	K-means	EM	通信業	K-means	EM
セゾン	90.1	88.9	NTT	84.8	90.9
オリックス	88.9	84.8	KDDI	86.9	78.8
三菱UFJFG	90.9	93.9	NTTドコモ	80.8	85.9
三井住友FG	96.0	90.9	ソフトバンク	93.9	89.9
みずほFG	92.9	83.8			

3.3 銘柄間での精度評価

3.1 節で生成した訓練データを用いて, 各銘柄同士でルールの生成と売買判断の予測を行った結果を評価する．ルール生成のための学習アルゴリズムは Weka に実装された PART[8]を利用した.

k-means および EM アルゴリズムによるクラスタリングによって時系列パターンを抽出したデータでの各銘柄間の売買判断を予測した正答率を表 4, 表 5 に示す．表側が訓練データとした銘柄，表頭がテストデータとした銘柄を表している．これらの結果から，訓練データの銘柄とテストデータの銘柄の業種が異なる場合でも正答率が高くなる銘柄があることが分かった．逆に，同業種でも正答率が低い組み合わせが存在した.

表 4 k-means によって時系列パターンを抽出した訓練データによる正答率(%)

K-Means	セゾン	三菱UFJFG	三井住友FG	みずほFG	オリックス	NTT	KDDI	NTTドコモ	ソフトバンク
セゾン		44.4	28.3	31.3	40.4	29.3	35.4	22.2	49.5
三菱UFJFG	46.5		44.4	30.3	42.4	32.3	39.4	29.3	**55.6**
三井住友FG	44.4	24.2		38.4	31.3	28.3	27.3	29.3	22.2
みずほFG	46.5	31.3	33.3		29.3	22.2	20.2	22.2	**58.6**
オリックス	38.4	**50.5**	27.3	31.3		32.3	39.4	19.2	30.3
NTT	14.1	**50.5**	27.3	31.3	14.1		39.4	37.4	6.1
KDDI	12.1	44.4	**56.6**	27.3	31.3	41.4		**55.6**	16.2
NTTドコモ	26.3	40.4	**52.5**	33.3	23.2	30.3	20.2		8.1
ソフトバンク	44.4	28.3	18.2	45.5	34.3	40.4	30.3	26.3	

表 5 EM アルゴリズムによるクラスタリングによって時系列パターンを抽出した訓練データによる正答率(%)

EM	セゾン	三菱UFJFG	三井住友FG	みずほFG	オリックス	NTT	KDDI	NTTドコモ	ソフトバンク
セゾン		46.5	28.3	31.3	38.4	**51.5**	**65.7**	21.2	32.3
三菱UFJFG	31.3		**51.5**	31.3	38.4	29.3	41.4	22.2	46.5
三井住友FG	23.2	**58.6**		34.3	31.3	43.4	32.3	30.3	29.3
みずほFG	35.4	31.3	34.3		31.3	42.4	38.4	43.4	20.2
オリックス	41.4	29.3	39.4	34.3		37.4	21.2	28.3	25.3
NTT	41.4	21.2	20.2	42.4	44.4		33.3	23.2	39.4
KDDI	**61.6**	**59.6**	**50.5**	28.3	27.3	42.4		28.3	37.4
NTTドコモ	27.3	42.4	29.3	**52.5**	25.3	30.3	19.2		28.3
ソフトバンク	**52.5**	45.5	27.3	31.3	41.4	33.3	43.4	19.2	

4. 考察

3 節における実験では，銘柄の組み合わせによって正答率に大きな差が生じた．この要因について，株価の絶対値の観点から考察する．Das らの方法，Keogh による Finding Motif などは株価の推移を波形として規則性を得ることを目的としているため，正規化によって絶対値の情報を除いている．しかし，本研究では絶対値の情報を含めた知識を提供するため，株価データなどは絶対値の形で利用した．図 4 は銘柄間での正答率と平均終値の差をプロットしたものである．

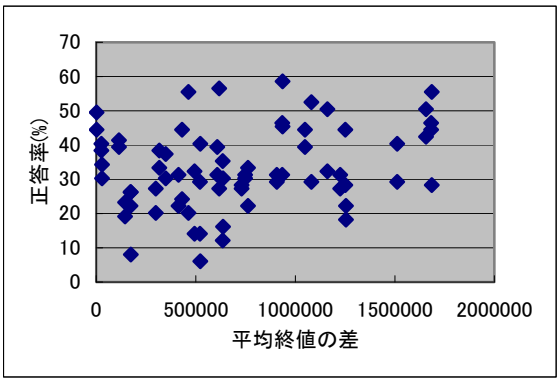

図 4 銘柄間での平均終値の差と正答率

図 4 およびこれらの間の相関係数が 0.26 であることから，株価の絶対値が必ずしも正答率に影響を与えているとはが言えないことが分かる．このことから，テクニカル指標による値と株価の時系列パターンの組み合わせによるルールでは，絶対値が重要となる場合と正規化された値が重要となる場合があると考えられる．各銘柄で売買判断に必要な指標の組み合わせはルールの条件部に記述されるため，正規化した値についても訓練データに追加していく必要があると考える．

また，正答率は勝率に当たるが，実際には売買タイミングを見極めることが利益を得るためには重要な要素となることが広く知られている．特に騰落レシオはその用途での有用性が指摘されており，このような市場全体のトレンドを反映する指標もデータに取り込んでいく必要があると考えられる．

5. おわりに

本稿では，株価データに対し，時系列ルール生成ツールを適用した結果について述べた．銀行・金融業 5 銘柄，通信業 4 銘柄について，20 日後に 5%の上昇／下降を売買の判断基準として，75 日間の株価およびテクニカル指標の時系列パターンから成る分類ルール集合の作成を行った．PART によるルール

集合によって各銘柄間での売買判断を予測した正答率は，各テストデータの最多クラス（判断結果）の割合を上回る銘柄の組み合わせがあることが判明した．

また，銘柄間の平均終値の差と正答率との間に強い相関が無いことが示されたが，より的確な売買判断ルールを作成していくため，今後，データセットの属性として絶対値の情報を取り除いた値を加えるなど，さらに評価を行っていく必要がある．

今後は，本ツールを基に株売買シミュレーションの結果をルール集合およびルール個々に示す機能の開発を進める．また，より的確な売買タイミングを取得するため，市場全体の値動きに関する指標の導入も検討していく．

謝辞

本研究の一部は科学研究費補助金基盤研究(B)「オントロジーとデータマイニングの統合に基づく知識マネジメント支援システム」(18300054)の助成による．

参考文献

[1] Raymond, W., and Ada, F., "Mining top-K frequent itemsets from data streams", Data Mining and Knowledge Discovery, 13(2) (2006) pp.193-217

[2] Lin, J., Keogh, E., Lonardi, S., and Patel, P., “Finding Motifs in Time Series”, in Proc. of Workshop on Temporal Data Mining (2002) pp.53-68

[3] Das, G., King-Ip, L., Heikki, M., Renganathan, G., and Smyth, P., “Rule Discovery from Time Series”, in Proc. of International Conference on Knowledge Discovery and Data Mining (1998) pp.16-22

[4] KabuRobo: [http://www.kaburobo.jp]

[5] Quinlan, J. R., “Programs for Machine Learning”, Morgan Kaufmann (1992)

[6] Quinlan, J. R., "Bagging, Boosting and C4.5", AAAI/IAAI, 1 (1996) pp. 725-730

[7] Witten, I. H. and Frank, E., “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”, Morgan Kaufmann, San Francisco (2000)

[8] Frank, E., Witten, I. H., “Generating accurate rule sets without global optimization”, in Proc. of the Fifteenth International Conference on Machine Learning (1998) pp.144-151