3A1-2

ヒントの改良にもとづくデータマイニングシステム

Data Mining System for Time-series Data based on Hints

杉村 博　　松本 一教
Hiroshi SUGIMURA　　Kazunori MATSUMOTO

神奈川工科大学大学院 情報工学専攻
Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology

This paper proposes a data mining system that discovers knowledge from time-series data based on hints. A user inputs hints, and the system uses these as suggestive patterns for time-series data analysis. The system predicts future situations by using decision tree based on similarities between hints and time-series data. In this method, the system can deal with user's knowledge. Furthermore, the system has a mechanism to improve hints using genetic algorithm. By this mechanism, even if a user inputs bad hints, we can obtain good knowledge. This paper describes details on the system and results of the experiment.

1. はじめに

時系列データとは時間とともに記録されたデータで，実社会での様々な場面で蓄積されている．時系列データマイニングとはこのような蓄積された時系列データから知識を発掘することを目的としている．本研究は 1 つの長大な時系列データから，そのデータの未来状態を予測する知識の発掘を行う．

本研究は，未来状態の予測のために指標となるパタンを用意して，そのパタンと見比べることで未来状態を予測するという人間的な行動を基にしている．このような手法の代表として，株価予測のテクニカル分析にダブルトップやダブルボトムがある．本研究ではこのような指標となるパタンの事をヒントと呼び，このヒントを用いることで人間の知識を基にした発掘を行うことのできるマイニングシステムを開発した．さらに，このシステムではヒントを自動的に改良することで，より良い知識を得る機能についても開発した．本論文ではこのシステムについての詳細と実験結果について述べる．

2. ヒントを用いた時系列データマイニング

本研究では時系列データの未来状態を予測する知識として決定木を作成する．しかし，従来の決定木学習手法は時系列属性を想定していないため，時系列データを含むデータ集合に適用する場合にはデータの前処理が必要となる．最も単純な方法の 1 つとして，時系列データを計測値の平均値で置き換える方法が考えられるが，この方法では時系列データの形を無視しており，たとえば形が大きく異なる時系列データが類似しているとみなされてしまう欠点がある．[山田 03]

そこで本論文では，人間の知識をもとにパタンを入力し，そのパタンを基に時系列データを予測するための知識を発掘するシステムを提案する．本研究では，このように時系列データを解析するための指標となるパタンをヒントと呼ぶ．ヒントは数値変化を定性的に表現したシーケンスであり，このヒントを基にして時系列データを解析することによって，時系列データの形を陽に扱うことができる．また，ヒントを定性的に表現することによって，様々なデータに対して汎用的に適用でき，株価データのように基準となる数値に一貫性がないデータに対しても適用できる．さらに，提案するシステムではヒントを自動改良し，さらに良い知識を発掘する仕組みも併せ持つ．この仕組みによって，解析するデータに対して知識のないユーザによって入力されたヒントでも，自動的にデータ解析を行うための指標となるパタンとなり，その改良されたヒントから発掘された知識は一定以上の精度品質となる．提案するシステムは，発掘する知識として時系列データの未来を予測するための決定木を作成する．

システムは長大な時系列データからいくつもの部分時系列データを切り出し，それら部分時系列データと各ヒントとの相違度を求めて教師データの属性値とする．ヒントはユーザによって入力された概念的な数値変化を示したデータであるため，実際のデータとのマッチングでは曖昧さを扱う必要がある．この方法については後述する．さらに，部分時系列データを未来状態によって分類し，その未来状態を教師データのクラスとする．未来状態は切り出した部分時系列データの最後の値を基準に決定する．ユーザはシステムに停滞幅倍率 x を与える．基準となるデータに対して $1+x$ 倍を超えたデータが未来状態に存在する場合に up に分類する．分類 up ではなく，かつ基準となるデータに対して $1-x$ 倍を超えたデータが未来状態に存在する場合に down に分類する．そして両方の条件を満たさない場合に分類を stay とする．このようにして作成した教師データの概要は図 1 のようになる．この方法によって作成した教師データを基にして未来状態を分類するための決定木学習を行う．

data	P0	P1	P2	class
0--19	0	103	100	up
15--34	33	15	98	down
20--39	78	15	153	up
25--44	78	15	153	stay
30--49	78	15	128	stay

図 1: ヒントを用いた教師データ

決定木学習は教師データを分割するための枝を生成する．生成した枝によって教師データを分割した際の評価値を計算し，その中の最大の評価値をとる分割によって教師データを分割

連絡先: 杉村 博，神奈川工科大学 情報工学専攻 松本研究室，神奈川県厚木市下荻野 1030，hiroshi.sugimura@gmail.com

する．評価値の計算方法として利得比基準 (gain ratio) を用いる [稲積 00]．また，各属性は連続値であるためその分割点 split info(X) は次の式で計算する．

$$\text{split info}(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|} \right)$$

分割方法が決定された後，その分割に基づいてデータは分割される．このようにして決定木学習はクラスを予測する知識を木構造によって表現する．

ヒントとの曖昧なマッチングを行うために動的計画法 (DTW) を用いる．DTW とは時間軸の非線形な伸縮を許し，コストという考え方によって 2 パタン間の距離を計算する手法である．時間軸のずれをコスト q, r，値の違いのコストを s とした時に，2 パタン間の距離は次の式で計算できる [Berndt 94]．

$$g(i,j) = \min\{g(i,j-1)+q, g(i-1,j)+r, g(i-1,j-1)+s\}$$

3. ヒントの改良

本システムはユーザによって概念的に与えられたヒントを，実際の時系列データの解析結果にしたがって改良することで，発掘する知識をさらに良いものへと改善する．さらに，この機能を用いることによって事前に知識を持たないユーザに対してはヒント自体を知識として得ることもできる．このために，本システムでは遺伝的アルゴリズム [Holland 75] を用いてヒントの改良を行う．

3.1 ヒントの遺伝子表現と適応度

本システムの遺伝子はヒントである．遺伝的アルゴリズムでは，この遺伝子を増加，減少させるために適応度を計算する．適応度は，各個体が解析する時系列データ中にどれだけ類似のデータが存在するかを表す．これは，人間がヒントと解析する時系列データを見比べた時に直感的な判断をしやすくするためである．ただし，すべての時系列データに対して類似してしまうと，分類を行う際の指標とはならない．このため，単純にヒントと部分時系列データとの相違度の平均値や中央値で計算できない．そこで本研究では決定木に用いられた枝の分割点を用いる．決定木学習は枝刈りのために信頼要因を設けて，信頼要因以上の条件付き情報量にならない枝を枝刈りする．このため，枝に設定された相違度の分割点には信頼要因以上の条件付き情報量が保障される．本システムではこの性質を利用して，枝に設定された分割点を遺伝子の適応度とする．

3.2 GA オペレータ

選択では基本的な手法であるルーレット選択を行う．この方法は遺伝子の適合度に応じた確率で次世代に残す遺伝子と削除する遺伝子を決定する手法である．ルーレット選択には適合度の低い遺伝子も残る可能性があり，局所解を防ぐ効果がある．

交叉では 2 点交叉を行う．この方法は交叉点を 2 つランダムで選択し，その交叉点で交叉する方法である．交叉によって適合度の高い遺伝子から，さらに適合度の高い遺伝子を作成できる可能性がある．

突然変異ではランダムに選択した遺伝子のシーケンスから，ランダムに 1ヵ所選択し，その値に対して乱数を増減する操作を行う．遺伝的アルゴリズムではこのような操作を行うことによって局所解に陥ることを防ぐ．

4. 株価データによる実験

1990 年 1 月 4 日から 2008 年 12 月 30 日までの実際の東証株式市場から合計 25 社の株価の過去データを使用してヒントを用いて決定木学習を行い，ヒントを改良しない場合と，ヒントを改良した場合の精度を比較する．ヒントの改良は，ヒントを 100 世代分改良したあとの決定木を用いる．

初期ヒントは 10 個とし，1 つのヒントに対して 10 個の乱数を発生させて作成した．スライドウィンドウのサイズは 20，スライド量を 5 とする．このとき，データの末端のスライドウィンドウのサイズに満たないデータは切り捨てる．未来状態は up，down，stay の 3 種類を用意し，クラスの分類数が均等になるような停滞幅倍率を予備実験によって求め，未来状態の観測期間を 5，停滞幅倍率を 5%とした．決定木学習アルゴリズムは C4.5 を使用し，決定木の分類精度は交差検定によって測定した．

GA before は本研究で提案した手法であるヒントに基づいた決定木である．GA after は遺伝的アルゴリズムによってヒントを改良したあとに出力した決定木である．Improve は GA before と GA after との精度の差である．

表 1: 株価変動による実験結果

	精度 (%)	time(s)	サイズ
GA before	50.4	7.9	72.1
GA after	63.4	950.8	41.7
Improve	13.0		

5. おわりに

ランダムで予想した場合の精度の理論値は 33.3%である．このことから，最も低い精度の株価データによる結果においても 50.4%の予測精度となっており，本研究によって手に入れた決定木は 1 つの指標として使える知識といえる．また，初期精度が低いものほど，遺伝的アルゴリズムによって精度が改善されていることがわかることから，ヒントの改良による効果は一定以上の精度品質を保つ効果があると考えられる．

ユーザの知識としてヒントを入力し，そのヒントを基にしてデータマイニングを行うシステムを作成した．さらに，ヒントを改良し，分類精度を向上する機能についても提案した．本システムを用いることによって，時系列データを解析する際の新しい指標を得ることができると考えている．

参考文献

[Berndt 94] Berndt, D. J. and Clifford, J.: Using Dynamic Time Warping to Find Patterns in Time Series, in *Proceedings of KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, pp. 359–370, Seattle, Washington (1994)

[Holland 75] Holland, J. H.: *Adaptation in Natural and Artificial Systems*, University of Michigan Press (1975)

[山田 03] 山田 悠, 鈴木 英之進, 横井 英人, 高林 克日己 : 動的時間伸縮法に基づく時系列データからの決定木学習, *IPSJ SIG Notes. ICS*, Vol. 2003, No. 30, pp. 141–146 (2003)

[稲積 00] 稲積 宏誠, 吉澤 有美 : 論理最小化に基づく決定木による知識発見, 人工知能学会誌, Vol. 15, No. 4, pp. 657–664 (2000)