ユーザ入力にもとづいた時系列データマイニングシステム

Mining system for Time-series Data with User Input

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 method of data mining based on decision tree for time-series data. It is important in practical areas to extract hidden knowledge by analysing time-series data. In particular knowledge that predicts future is worth much in most cases. The decision tree approach is applicable to this purpose, however, simple applications often cause problems. We in this study a pattern, which is given by a user, is used as a hint of knowledge discovery. We focus the scope of discovery on relating portions to the given pattern. As a result, the resulting decision tree becomes simple and easy to understand. Then we can convent useful and effective knowledge from the decision tree. We explain an outline of the system, and investigate experimental results.

1. はじめに

データマイニングの重要な技術の 1 つに決定木がある．決定木から容易に IF-THEN ルールを抽出できる利点もあり，成功事例も多数報告されている [杉井 07]．しかし時系列データに対する適用は少ない．単純な方法では決定木のサイズも大きくなり，分類精度の悪い決定木となる．そこで本研究ではユーザ入力によりデータ中の着眼点をヒントとして与えることで，単純で分類精度の高い決定木を作成する方法を開発した．本論文ではシステムの詳細と，実際に学習した決定木を用いた実験結果，そしてユーザ入力を用いずに作成した決定木との比較を行う．

2. 時系列データマイニング

時系列データとは毎日の気温や株価，視聴率のような時間の経過に沿って記録したデータのことであり，実社会のさまざまな分野で頻出し重要視されている．一方データマイニングとは，収集したデータから役立つ情報を発見する技術である．データマイニングを実現する技術には様々な方法がある．相関ルールの抽出にもとづく方法も研究されている [杉村 08]．本論文では決定木を用いる方法について提案する．

2.1 決定木によるデータマイニング

決定木によるデータマイニングでは，各々のデータがクラス属性を持ち，その値（クラス値）が与えられるものとする．また，データは定められえた個数の属性の値（属性値）の組として記述される．クラス値を判定するための，属性値のテストを行うことのできる，もっともコンパクトな決定木を得ることが目標となる．

時系列データは各時点での値の組として与えられるが，それらを単純に属性値とみなして決定木を作成した場合，データの特徴を十分表現することができず，得られる決定木は複雑で精度の低いものとなってしまう．そこで本手法では，以下に述べるユーザ入力により，時系列データの注目点を絞り，そこから属性（および属性値）を抽出するようにした．

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

2.2 ヒントとしてのユーザ入力

ユーザによる手書き入力グラフをヒントとして用いる．ユーザはポインティングデバイスを用いてグラフを記述する．このグラフをサンプリングするように等間隔に数値化する．グラフの分割数が大きいほど詳細に数値化し，分割数が小さいほど荒く数値化する．図 1 に実行例を示す．

ユーザはシステムに手書き入力による時系列データ中の着眼点として，グラフの特徴を与える．本システムはこのユーザ入力に類似した部分時系列データをもとにして決定木を作成する．このように，類似した部分時系列データのみを抽出して，学習することにより特定条件において精度の高い決定木を作成できる．

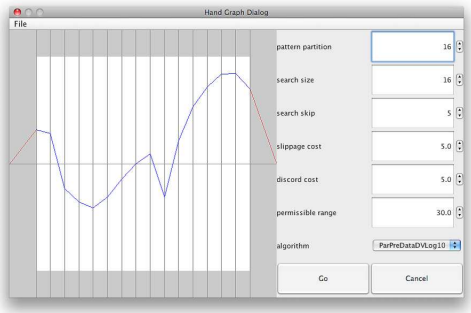

図 1: 手書き入力ダイアログ

ユーザ入力は手書き入力グラフだけでなく，グラフ分割数，部分時系列データの区切りサイズとスキップ幅，DP マッチング時のずれコスト，不一致ペナルティ，許容類似度を入力する．これらの内，グラフ分割数以外の値は次節で使用するパラメータとなる．

2.3 DTW によるマッチング

DTW(Dynamic Time Warping) とは，パターンの要素間に定義された類似度にもとづいて，パターンの伸縮まで考慮に入れたマッチング方式である [Last 04, 大桃 05]．マッチングの際に必要なプロパティは 3 つあり，それぞれ，ずれコスト，

不一致コスト，許容類似度である．ずれコストを q，r，不一致コストを s とした数式は式 1 となる．

$$g(i,j) = \min \begin{cases} g(i,j-1)+q \\ g(i-1,j)+r \\ g(i-1,j-1)+d(i,j)+s \end{cases} \quad (1)$$

3. 株価データによる実験

本論文では典型的な時系列データとして，株価のデータによる実験を行った．図 2 に実験の概要を示す．

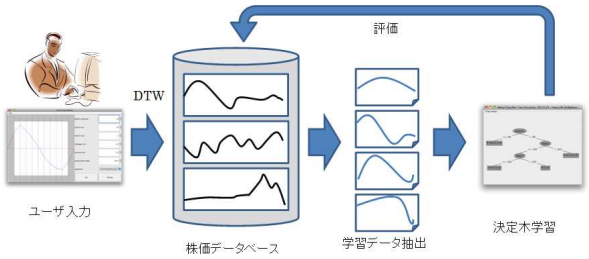

図 2: 実験概要

本論文にて作成したシステムを用いて得られたデータを教師データとして，データマイニングツール Weka[Waikato 09] を用いてクラス分類を行った．用いた手法は決定木学習で，使用したアルゴリズムは C4.5 である．

決定木は，買い評価用の決定木と売り評価用決定木の 2 種類を用意した．決定木のクラスは 2 値とし，買い評価用の決定木は「買いの場合」には yes,「買いではない場合」を no とした．売り評価用の決定木も同様に,「売りの場合」には yes,「売りではない場合」を no とした．

「買いの場合」とは短期的に上昇する未来予測であり，検索結果の時系列データの初期値よりも高い値が，検索結果最終日から 5 日以内に 1 つ以上存在していた場合と定義した．同様に「売りの場合」とは短期的に下降する未来予測であり，検索結果の時系列データの初期値よりも低い値が，検索結果最終日から 5 日以内に 1 つ以上存在していた場合と定義した．表 1 が買い評価用決定木の結果データ例，表 2 が売り評価決定木の結果データ例である．

さらに,「買いの場合」についてはユーザ入力による関係性が高いデータの選抜を行わずに，すべての部分時系列データを教師データとして決定木を作成し，ユーザ入力の有無による差を表 3 にまとめた．

series は部分時系列データの総数，hit は手書き入力にマッチした部分時系列データ数，hit%は部分時系列データの総数に対する割合，Sott は作成した決定木のサイズ，CCI は作成した決定木の分類精度である．

4. おわりに

本研究では時系列データから未来予測を行う決定木を作成するシステムを開発した．このシステムではユーザが手書き入力によるヒントを与えることによって，類似の部分時系列データのみを得ることができた．類似の部分時系列データから得られる決定木は，ユーザ入力が無い時と比べて単純化し，視覚的に理解しやすい決定木となった．さらにユーザ入力が無い時と比べて分類精度も向上した．

表 1: 買い評価結果例

stock No	series	hit	hit(%)	Sott	CCI(%)
1	329	103	31.3	15	72.8
2	329	136	41.3	5	63.4
6	329	127	38.6	33	58.3
13	329	132	40.1	7	90.9
17	317	118	37.2	17	69.5
平均	326	121	37.1	12	75.5

表 2: 売り評価結果例

stock No	series	hit	hit(%)	Sott	CCI(%)
1	329	108	32.8	15	50.0
2	329	94	28.6	19	67.0
6	329	104	31.6	9	48.1
13	329	105	31.9	1	51.4
18	317	86	27.1	1	46.4
平均	326	105	32.1	7	56.7

表 3: ユーザ入力の有無の差

	入力有		入力無		入力有と無の差	
stock	Sott	CCI	Sott	CCI	Sott(%)	CCI(%)
1	15	72.8	93	63.3	15.8	9.5
2	5	63.4	53	70.6	9.4	-7.2
6	33	58.3	73	54.8	45.2	3.5
13	7	90.9	83	61.5	8.4	29.4
17	17	69.5	7	64.8	242.9	4.7
平均	12	75.5	65	63.0	18.8	12.8

ここで得られた相関ルールを用いて，実際の株価データによる株の売買シミュレーションを行うことでさらに問題点や改良を行うことが今後の研究課題である．

参考文献

[Last 04] Last, M., Kandel, A., and Bunke, H.: *Data Mining In Time Series Databases*, World Scientific (2004)

[大桃 05] 大桃 論, 陳 漢雄, 古瀬 一隆, 大保 信夫 : タイムワーピングに基づく時系列データの類似検索 - 次元縮小による効率化, *DBSJ Letters*, Vol. 4, No. 1, pp. 17–20 (2005)

[杉井 07] 杉井 学, 松野 浩嗣 : 機械学習によるスパムメールの特徴の決定木表現, *IPSJ SIG Notes*, Vol. 2007, No. 16, pp. 183–188 (2007)

[杉村 08] 杉村 博, 松本 一教 : DP マッチングを利用する時系列データからのデータマイニング, 第 22 回 人工知能学会全国大会論文集 (2008)

[Waikato 09] Waikato, of T. U.: Machine Learning Project at the University of Waikato in New Zealand, http://www.cs.waikato.ac.nz/ml/ (2009)