F-030

新しい時系列データクエリの記述方法をもちいた知識発見
Knowledge discovery by using new time-series data base query description

杉村 博　　松本 一教
Hiroshi Sugimura　Kazunori Matsumoto

1. はじめに

本論文は時系列データの未来状態を予測するための知識抽出を目的に，2 つの特徴をもつシステムを提案する．1 つ目のポイントは時系列データクエリ言語の提案である．時系列データの未来状態を予測する IF-THEN スタイルの知識を抽出するために，機械学習を用いることができるが，与えるデータに依存して得られる知識の有用性が異なる．このため，与えるデータの選択は重要な役割がある．データ選択の方法として類似時系列データを検索し，収集する方法[4]があるが，1 つのシーケンスとの類似度によって検索を行うだけでは柔軟性に乏しい．そこで複数のシーケンスの組み合わせの記述によって検索をおこなう時系列データクエリ言語について提案する．

2 つ目のポイントは未来状態の予測を行うための特徴的なパタンの自動的な発見と改良である．決定木学習において時系列データの特徴を属性として扱うことが重要であるため，相違度を属性値とする方法を提案している[6]．この手法では良い特徴的なパタンを与える必要があり，知識のない人間にとっては扱うことが難しい．そこで，この特徴を自動的に発見する手法について提案する．

2. 時系列データクエリ言語

この言語はユーザによって入力されたシーケンスとそれらの組み合わせによってクエリを記述する．検索するデータは時系列データを対象としているため，シーケンスの単純な組み合わせ表現だけでなく，それらの出現順序や出現距離も重要である．出現順序は XPath のように "/"を用いてつなげる経路表現として記述し，出現距離はワイルドカードを用いて正規表現のように記述する．表 1 にクエリの構文を，表 2 にワイルドカード表現を示す．

表 1. 時系列データクエリの構文

Query	::=	Path | Path Op Query
Path	::=	Pattern | Pattern "/" Path
Pattern	::=	String | Wildcard
Op	::=	"and" | "or"
Wildcard	::=	表 2 を参照

表 2. ワイルドカード表現

.	何か 1 つのデータにマッチ
*	0 以上の連続にマッチ
+	1 以上の連続にマッチ
?	0 か 1 個の連続にマッチ
{n, m}	n 個以上 m 個以下の連続データにマッチ

このクエリを用いることでシーケンスの出現順序と出現距離を定義して柔軟な検索が可能となる．たとえば交通量のピークとピークの中間部分にはどのような頻出シーケンスとそれらの関係があるかの調査や，異なる銘柄の株価データから買いを示唆するシーケンスと売りを示唆するシーケンスの出現距離を指定して検索を行うことができる．図 1 にシーケンスを定義してクエリを記述し，検索を行った結果の例を示す．例示したクエリは「P0 に類似したデータが出現した後，P1 と類似したデータが出現するデータを検索する，このとき 50 個以下ならばその間にどのようなデータが出現してもよい」ということを意味している．クエリに記述した P0, P1 と比較している数値はシーケンスと検索データとのマッチを行うための相違度を計算した時に，類似していると認識するための閾値である．相違度の計算方法は次節で説明する．

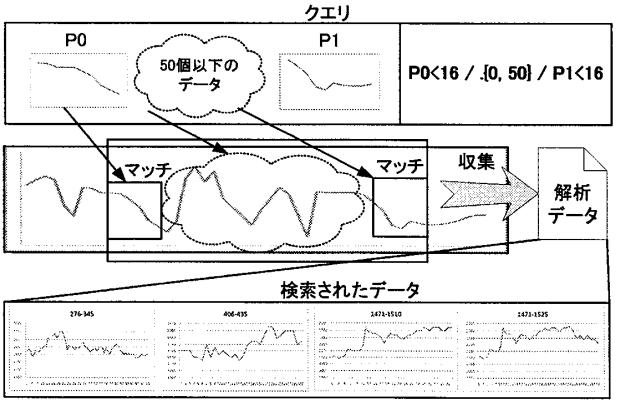

図 1. クエリと検索結果の一部

3. Dynamic Time Warping

シーケンスと時系列データとの相違度の計算は Dynamic Time Warping(DTW)によって行う．DTW は，パタンの要素間に定義された類似度に基づいて，パタンの伸縮まで考慮に入れたマッチング方式である[2]．マッチングの際に必要なプロパティは時間軸のずれに対応したコストと，値の一致度に対応したコストである．計算した距離がクエリ言語によって入力された閾値以下のときに，パタンにマッチした個所となる．

4. 知識抽出

本研究では決定木学習を用いて知識の抽出を行う．決定木学習で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する[1]．一定の基準を満たすまで再帰的に行い，クラスを予測する知識を木構造によって表現する．

このアルゴリズムを用いて本研究では時系列データの特徴から未来状態を予測する決定木を作成する．ただし，単純な決定木学習では時系列データを属性として扱うことができないため，属性には時系列データと特徴を表すシーケンスとの相違度によって表現する．この相違度の

神奈川工科大学大学院 工学研究科情報工学専攻
Kanagawa Institute of Technology, Course of Information and Computer Sciences.

計算には上述した DTW を用いる．そして未来状態によってクラスを付与する．作成したトレーニングデータの例を図 2 に示す．

P0	P1	P2	P3	class
497.03	4636.99	924.04	1190.33	up
457.12	4313.78	884.13	1150.42	up
212.08	2032.27	626.83	905.15	down

図 2. トレーニングデータの例

4.1 特徴の発見と改良

決定木の予測精度は属性として与える特徴に依存する．特徴は人間が与えることで高い予測精度の決定木を作成できるが[6]，この方法では知識を持たないユーザにはよい特徴を与えることができない．そこで本研究では属性として与える特徴を自動的に発見し，改良を行うことによって高い予測精度の決定木を作成する．この特徴の発見と改良のメカニズムには遺伝的アルゴリズムを用いる．

まず，乱数を用いて遺伝子集団を生成する．各遺伝子の適合度を計算し，選択，交叉，突然変異のオペレータによって遺伝子集団を改良する．図 3 にオペレータの動作概容を示す．システムは終了条件を満たすまでこの操作を実行する．

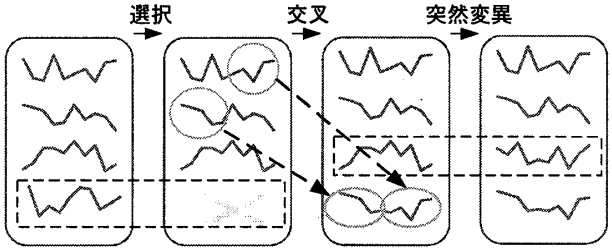

図 3. オペレータの動作概容

遺伝子の評価値には，特徴として未来状態を分類する際に用いた場合の獲得情報量[3]を用いる．

5. 実装と実験

実際にシステムを作成し，マイニングを行う．作成したシステムのアルゴリズムを図 4 に示す．ユーザはシステムにクエリを入力する．システムはクエリに従い時系列データを検索し，収集する．収集したデータを前データとの比にすることで標準化する．標準化したデータに分類を付与してトレーニングデータを作成する．このトレーニングデータに対して特徴の発見と改良を行いながらパタンを抽出する．最後に，抽出したパタンの評価を行い，評価結果とともに出力する．

データは 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 10 社の株価の過去データを使用する．未来状態は up, down, そして stay の 3 種類によって分類する．

従来の手法によって類似データを収集して未来状態の予測を行った場合と，提案したクエリを用いて収集したデータから未来状態の予測を行った場合とを比較した結果を表 3 に示す．表 4 には遺伝的アルゴリズムによる特徴候補の改良前と後の予測精度の差を示す．

Algorithm: Datamining algorithm

```
input: time-series database DB
output: the set of all feature patterns in elite population and
decision tree
begin
  input query;
  seach sequences in DB based on query;
  normalize sequences;
  classify sequences;
  // start genetic algorithm
  make a initial population which is a set of genes;
  elite = 0;
  for g = 1 to termination condition
    evaluate fitness of population;
    select genes in population;
    crossover genes in population;
    mutation genes in population;
    if now population is best population then
      tree = DecisionTreeLearning( population );
      elite = g
    end;
  end;
output population_elite and tree;
end;
```

図 4. 実験用アルゴリズム

表 3. 単純な類似検索との比較

	単純な類似検索	時系列データクエリ
検索数	212	103
予測精度	40.09%	48.54%

表 4. 遺伝的アルゴリズムによる予測精度の上昇

	改良前	改良後	上昇量
平均予測精度	50.43%	54.90%	4.47%

6. おわりに

1 つのシーケンスとの類似時系列データを収集して未来状態を予測するよりも，いくつかのシーケンスの組み合わせによってデータを検索するほうが，適切なデータを獲得できることが明らかになった．また，特徴を自動的に発見，改良することによって知識のないユーザにとっても有用な未来状態を予測する知識を獲得できることも明確になった．

参考文献

[1]Chia Hui Huang, Wen Ching Liou, Berlin Wu, "A Decision Support System for Stock Investment of Listed Companies in Taiwan", Biomedical Soft Computing and Human Sciences, Vol.13, No.1, PP.31-36 (2008).

[2]Eamonn J. Keogh, Michael J. Pazzani, "Derivative Dynamic Time Warping", In First SIAM International Conference on Data Mining(SDM'2001), pp.359-370 (2001).

[3]Ian H. Witten, Eibe Frank, "DATA MINING: Practical Machine Learning Tools and Techniques", Morgan Kaufmann Pub (2005).

[4]Martin Wattenberg, "Sketching a graph to query a time-series database", CHI '01: CHI '01 extended abstracts on Human factors in computing systems, pp.381-382 (2001).

[5]Saroj Kaushik, Naman Singhal, "Pattern Prediction in Stock Market", AI 2009: Advances in Artificial Intelligence, Vol.5866, pp. 81-90 (2009).

[6]杉村 博, 松本 一教, "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", 第 23 回 人工知能学会 全国大会論文集 (2009).