1F2-01

多数の欠損値を持つ時系列データからのデータマイニング手法の一検討

Data Mining from Time Series Data with Many Missing Values

本山 真也*1　市瀬 龍太郎*2　沼尾 正行*3
Shinya MOTOYAMA　Ryutaro ICHISE　Masayuki NUMAO

*1 東京工業大学 大学院 情報理工学研究科
Department of Computer Science, Tokyo Institute of Technology

*2 国立情報学研究所 知能システム研究系
Intelligent Systems Research Division, National Institute of Informatics

*3 大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

This paper proposed the data mining method based on common periods to create a rule set from time series data with many missing data and irregular intervals. For the reason mentioned above, this mining method must have the function to accommodate these features. This method is the attempt to focus on some periods when common behavior is shown among some data and to combine them to create rules.

In order to evaluate the proposed method, the authors conducted experiments with medical time series data.

1. はじめに

近年，自然科学の分野だけでなく，医学，経済学などの分野においても，時間とともに不規則に変動するデータを扱うことが増加してきており，こうした時系列データに対するデータマイニングの重要性が増大している．目標は過去の情報の分析，モデルの構築，将来予測と様々であり，代表的な時系列データとしては，株価データや血液検査データがあげられる．

本研究ではこのような時系列データから有用な知識を発見する手法として，データが共通の振る舞いを示す期間に着目したデータマイニング手法を提案し，その性能について検討した．

2. 時系列データを扱う際の注意点

時系列データからデータマイニングを行う際の注意点として次の 2 点に注目した．

- 間隔が不定期である時系列データを，直接扱えるデータマイニング手法は少ない．
- 時系列データの欠損値に対して，何らかの処理をする必要がある．

2.1 時系列データの間隔が不定期である場合

データマイニング手法としてはクラスタリング，決定木，ニューラルネット，遺伝アルゴリズムなど様々なものが存在するが，これらの手法は間隔が一定の時系列データを扱う手法である．したがって，時系列データの間隔が不定期である場合にこれらの手法を使うためには，データを月ごとに平均化するなど，何らかの処理が必要となる．また，これらの手法は時系列データの属性間の関係が強い場合にも適していないという問題点がある．

2.2 時系列データの欠損値の取り扱い

従来のグラフの類似性を利用するデータマイニング手法は，時系列データが定期的に収集されていると仮定している．しかし，不定期に収集され，欠損値が多くなる傾向の強い時系列データに対しては，この仮定が成り立たないのでこれらの手法は有効に働かない．ある時点のデータを取得できなかった場合，グラフでは形状が異なって見えるが，その時点のデータを取得できたグラフと実際には同じ傾向である可能性がある．しかし，上述の手法では両グラフは類似していないと判断してしまうという問題がある．このように，欠損値が多い時系列データを扱う際にはこの点に配慮をする必要がある．

連絡先: 本山 真也, 東京工業大学 大学院 情報理工学研究科 計算工学専攻，〒 152-8552，東京都目黒区大岡山 2-12-1, moto@nm.cs.titech.ac.jp

3. 提案手法

前述のような問題点がある時系列データから有用な知識を発見する手法として，データが共通の振る舞いを示す期間に着目したデータマイニング手法を提案する．ここで，データが共通の振る舞いを示す期間とは，複数のデータについて，ある属性の属性値が同一である期間を意味するものとする．例えば，データ A とデータ B は時点 t_1 から時点 t_2 まで，常に属性値が 100~200 の範囲内にあったという形式のものとなる．

まず，この手法は時系列データの表形式データを入力とし，最終的に時系列データがある規則に照らして正しいか正しくないかを判定するルールを出力とする．ただし，ルールは分類する条件の連言で表すようにしている．

具体的な手順は**表 1** の通りである．

表 1 提案手法のアルゴリズム

入力：時系列データの表形式データ
出力：ある条件に照らして時系列データを正例と負例に分類する規則の集合
処理手順：
1. 正例のみから属性ごとに，属性値が同一の期間を求める．
2. 複数のデータでその期間が重なる部分を共通の振る舞いを示す期間とする．
3. 求めた期間から，負例が同様の振る舞いをするものに負の評価点をつける．
4. 最後に最小記述長原理により期間を組み合わせて，正例と負例を分類する規則を出力する．

3.1 提案手法の狙い

提案手法はまず正例のみから，共通の振る舞いをする期間を求めている．このことによりできるだけ多くの正例が共通の

表 2 各試験での正答率

データセット	1	2	3	4	5	6	7	8	9	10	平均
正答率 (%)	0.62	0.54	0.62	0.54	0.38	0.54	0.31	0.46	0.54	0.62	0.52

振る舞いを示す期間を求める事ができる．こうして正例を多く含み，負例をなるべく含まないような規則を見つけることにより，支持度や確信度を上げる効果を期待している．

また，月ごとにデータを平均化するなどの処理を行う必要がないため，時系列データを直接扱うことができる．

最後に欠損値の扱いであるが，時系列上の直前と直後のどちらかの値であるものとしている．近似したりせず，曖昧にしておくことで，共通の振る舞いを示す部分を幅広く求める狙いがある．

4. 提案手法の性能評価実験

提案手法の性能を検討するため，欠損値が多い時系列データとしてインターフェロン (IFN) 投与患者の血液検査データを用いて，IFN 治療が有効であるか無効であるかのルールを作成する実験を行った．ここで，有効とは HCV-RNA(肝炎ウィルスの有無の判定法) により IFN 投与後にウィルス消滅を確認したものとし，無効は IFN 投与後にウィルス存在を確認したものとする．

4.1 データの前処理

まず血液検査データより，IFN 投与 3 年間前までの検査データを抽出した．次に，検査項目を列とし，患者番号 (MID) および検査日をキーとして，各検査項目が個別の属性となるような表形式データに変換した．

検査項目については，特に重要であると思われる GOT, GPT, TTT, ZTT, D-BIL, I-BIL, T-BIL, ALB, CHE, TP, T-CHO を使用し，検査値を医師の作成した離散化指標を基に 4 ～ 7 段階に離散化した．

このように前処理した検査データの中で，IFN 投与後にウィルスが消滅した 55 例を正例，IFN 投与後にウィルスが残っていた 82 例を負例とした．

4.2 実験結果

前処理済みデータを用いて，10-fold の交差検定を行った．各試験における得られた規則集合の正答率は**表 2** の通りである．ここで，正答とは，正例が得られた規則に当てはまる場合，あるいは負例が得られた規則に当てはまらない場合とした．

ここで，実際に得られた規則 (**図 1** の楕円で囲まれた部分) を 1 つ紹介する．横軸は 0 を IFN 投与開始日とした時間軸 (単位：日) であり，縦軸は TP(総蛋白) の値 (単位：g/dl) である．

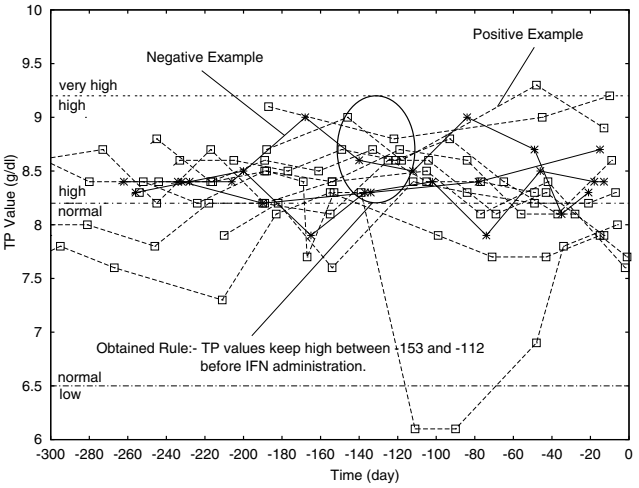
この規則には，正例 55 例中 12 例が該当し，負例 82 例中 3 名が該当した (確信度 80%，支持度 21.8%)．つまり，該当期間においては多くの負例の TP の値は HIGH ではなく，正例集合の中に規則に該当するグループが存在することが発見できた．

4.3 考察

図 1 のような高い支持度の規則が得られたことにより，できるだけ多くの正例を含む規則を見つける事で支持度を高める効果があることが確認できた．

また，平均の正答率が 52% であり，Progol による手法 [1] の正答率 52.6% と同程度の値が得られた．本研究では正例のみから共通の振る舞いを示す期間を求めた後に，負例により負の評価点をつけているため，負例の数が増えるほど得られた規則の正答率が悪くなる可能性が大きくなる．Progol による手法 [1] よりも負例の数が 2 倍以上多いにも関わらずほぼ同等の値を得ることができたので，同数の負例数であればより高い性能を発揮できるものと考えている．

図 1 得られた規則の例



IFN 投与 153 日前から IFN 投与 112 日前まで常に属性 TP の値が HIGH (8.2 以上，9.2 未満) である患者に対しては，IFN 投与が有効である

5. むすび

本論文では，欠損値の多い時系列データに対し，有用な知識を発見する手法としてデータが共通の振る舞いを示す期間に着目したデータマイニング手法を提案した．欠損値の多い時系列データとして実際の医療データを用いて実験を行った結果，高い支持度の規則が発見できることがわかった．また，正答率についても欠損値の多い時系列データに対して一定の効果があることが示せた．

今後は，正答率をさらに向上させるため，アルゴリズムの変更を行う．具体的には，現在は正例のみから共通の振る舞いを示す期間を求めているが，負例について考慮した形で生成する予定である．これにより，より多くの正例を含む規則を生成する過程で，負例を余計に含むケースを減らすことを目指している．

参考文献

[1] 佐藤慶宜，市瀬龍太郎，横井英人，沼尾正行，インターフェロンの効果を予測する述語記述の発見，人工知能学会研究会資料，SIG-KBS-A304-05，pp. 25-30, 2004.

[2] Ryutaro Ichise, Masayuki Numao, Discovery of Temporal Relationships using Graph Structures, Proc. of the 2nd International Workshop on Active Mining, pp. 118-129, 2003.