平成 15 年度　卒業論文概要

入学年度 平成 12 年	**学生番号** 12117742	**氏名** 水川 博之
卒業研究題目 論理的帰納法を用いた時系列医療データのマイニング手法		和田・犬塚 **研究室**

1. はじめに

データマイニングとは, 大規模なデータベースからあるパターンやルールを発見するプロセスである. 本論文では時系列医療データに注目し, 帰納論理プログラミング (Inductive logic Programming, **ILP**) によってデータの解析をおこなうため, そのためのデータマイニング手法について検討することを研究目的とする. そのために ILP を用いたデータマイニングについての形式およびその手続きを与え, さらに医療データの論理表現方法, その表現を利用した背景知識を提案することで, データマイニングをおこなう.

また, 医療の時系列データを扱った例 [1] は少ないが,ILP は述語論理を使用することで表現力が豊かであり, 背景知識として医療情報を組み込むこともできるので, 可読性の高い解析ができると考えられる.

2. ILP

ILP とは, 述語論理上で帰納推論を展開するアプローチで, さまざまな分類問題を解決することのできる枠組みのことである.ILP はこれまでにも分類学習のために用いられてきた [2]. 分類学習においては,ILP は次の式を満たすような仮説 H を求めることが目的である.

Given B:背景知識,EX^+:正事例集合,EX^-:負事例集合

$$\begin{cases} B \cup H \models ex^+, \forall ex^+ \subseteq EX^+ \\ B \cup H \not\models ex^-, \forall ex^- \subseteq EX^- \end{cases}$$

Find H:仮説

3. ILP を用いたデータマイニング

通常の ILP が仮説を一つ求めるのにたいして, データマイニングでは事例から観測されるルールを網羅的に見つけることが目的である. 本論文では ILP を用いたデータマイニングは次のように定式化する.

Given B:背景知識,EX^+:正事例集合,EX^-:負事例集合

$$\begin{cases} B \cup R \models ex^+, \forall ex^+ \subseteq X, X \subseteq EX^+ \\ B \cup R \not\models ex^-, \forall ex^- \subseteq EX^- \end{cases}$$

Find すべての R:仮説

この式を用いた手続きを次に示す.

1 Given B:背景知識,EX^+:正事例集合,EX^-:負事例集合
2 R:=ϕ
3 Repeat
　B'⊆B を適当に選択
　B',EX^+,EX^- に対して ILP を用いて
　仮説（=節 (ルール) 集合）H を導出
　R:=R∪H
4 End Repeat
5 R から正事例 EX^+ に対するサポートの高い
　ルール集合 R'⊆R を選択
6 Return R'

4. ILP を用いたデータマイニングへの適応

ILP では, 生データを論理表現することで解析される. 本論文では時系列医療データを用いる. 医療データには次のような特徴がある.

- 患者が検査をおこなうごとにデータは増加する
- データは一定間隔ではない
- データは欠損している

ここでの問題点は欠損値の扱いである. 決定木やニューラルネットなどの従来法では欠損値に対応できず,ILP を用いる理由がこれである. ではその論理表現方法とは,

$$testdata(TEST, ID, DATE, VALUE).$$

のように, 検査項目ごとに表現することで欠損値にも対応可能となる.

ILP で解析をおこなううえで, 背景知識としていくつか提案する. 背景知識を用いることによって次のような利点がある.

- 分かっている知識を仮説生成に使える
- 理解しやすい知識を与えることで, 仮説も理解しやすくなる
- 背景知識集合が仮説の探索空間をきめることになるため, 探索の制御に利用できる

つまりここで与える背景知識がデータマイニングをおこなううえで最も重要になる. そこで, 与える背景知識として次の 8 つを作成した.

1. average_qv(TEST,ID,QVALUE):検査値の平均をもとめて範囲 (high or low) を出力
2. maximum_qv(TEST,ID,QVALUE):検査値が最大となる日付近の検査値の範囲を出力
3. minimum_qv(TEST,ID,QVALUE):検査値が最小となる日付近の検査値の範囲を出力
4. all_qvalue(TEST,ID,QVALUE):検査値が high となる日が検査日の 3 割より多ければ high を,low となる日が多ければ low を出力
5. change(TEST,ID,CHANGE):検査日を 3 つに分割し, 検査値の平均が前半期間＜中盤期間＜後半期間ならば,increase を, 前半期間＞中盤期間＞後半期間ならば decrease を
6. variance_range(TEST,ID,QVALUE):検査値の分散を求めて, 他の患者と比較し大きければ big_variance を出力
7. outlier(TEST,ID,DATE,OTLIER):任意の日付において特異点の (他の検査値と比べて極端に高いまたは低い) 日を出力
8. frequency(TEST,ID,NUMBER):検査の回数の頻度 (多ければ big_number) を出力

5. 実験とまとめ

実験として, 正事例を thrombosis(血栓症) 患者, 負事例を正常患者として, 感染一年前からの検査データを入力することで感染患者のパターンを発見する. 実験には ILP システムとして代表的な Progol を用いた. 結果として評価の高かったもの 3 つを示す.

- maximum_qv(wbc,ID,high)⇒thrombosis(ID)
  検査 WBC において検査値が最大となる日付近では,high となる傾向がある患者は感染の疑いがある
- minimum_qv(tp,ID,low)⇒thrombosis(ID)
  検査 TP において検査値が最小となる日付近で,low となる傾向がある患者は感染の疑いがある
- change(wbc,ID,increase)⇒thrombosis(ID)
  検査 WBC において検査値が上昇傾向である患者は感染の疑いがある

本論文では,ILP を用いたデータマイニングのために, 形式と手続きを与えた. そして作成した背景知識を与えて実験をおこなったところ, 高い評価が得られたので,ILP を用いた時系列医療データのマイニングは可能であることが検証できた.

参考文献

[1] R.Ichise,M.Numao : Relational Mining for Temporal Medical Date, Proceeding of the LASTED International Contence in Intarmation and Knowledge Sharing, 2003.
[2] 古川康一, 尾崎知伸, 植野研 : 帰納論理プログラミング, 共立出版, 2001.