D-033

相関ルールを利用した時系列健康診断データの解析
Association rule is applied to time series of physical examination data

石川 亮一[1]　小尾 高史[2]　山口 雅浩[1]　大山 永昭[3]　佐々木 敏雄[4]
Ryoichi Ishikawa　Takashi Obi　Masahiro Yamaguchi　Nagaaki Ohyama　Toshio Sasaki

1．はじめに

近年健康診断の結果の電子化が進み、電子データとして記録されつつある。従来の健康診断データ利用法は健康診断受診時の値が正常か異常かを判断することに利用されてきた。しかし電子データとなることにより、これまでは困難であった検査項目の長期間の推移を見ることや、多くのデータを一度に取り扱うことが出来るようになった。そのため電子データを基にした解析が容易になりる。疾病に罹患するまでの過程を解明することで、健康状態の推移を明らかにし、生活習慣病の疾病予防に役立つことが期待できる。

そこで本研究の目的は、時系列健康診断データに相関ルールを利用し、時系列の波形とある疾病の間にどのような関連性があるのかを解明することで、健康状態を表現する手法を開発することを目指す。

2．解析の流れ

本研究では、毎年定期的に行われている健康診断の結果を数年にわたり記録されているものを利用する。今回の発表では疾病の対象を高血圧症とし、検査値の推移波形と高血圧症の関連性について相関ルールを適用して１年後に疾病に罹患する確率を求めた。以下に解析の手法を示す。

① 対象とする疾病の決定
② 健常者・疾病罹患者の分類（医師によって疾病ラベルがつけられている）
③ 短期間データへの投影
④ 時間軸空間への投影
⑤ 時間軸空間の分割(分割領域と呼ぶことにする)
⑥ 教師データに対して相関ルールを適用
⑦ Confidence（１年後に疾病に罹患する確率）を計算する
⑧ テストデータを教師データで求めた空間に適用し、教師データの confidence の妥当性を確認する

3．相関ルールについて

≪相関ルール≫の定義

Confidence：前提部の条件が起きるもとで結論部がどのくらいの確率で起きるかを示す。本研究では検査値がある時間的推移を示すデータ数が前提部にあたり、その時間的推移を示したデータのうち１年後に疾病ラベルが付けられたデータ数が結論部にあたる。

Support：前提部と結論部が同時におこる確率を示す。本研究では全体のデータ数のうち、ある分割領域内のラベルありのデータ数の割合を示す。

≪相関ルールの適用方法≫

① n 年分の時系列の波形を図１のように n 次元空間上に投影する
② 多次元空間を分割して、その分割した１つの領域を「分割領域」と呼び、図２の手順で分割する。
- Support が閾値以上であるならさらに分割を続ける
- Support が閾値未満ならそこで分割をやめる

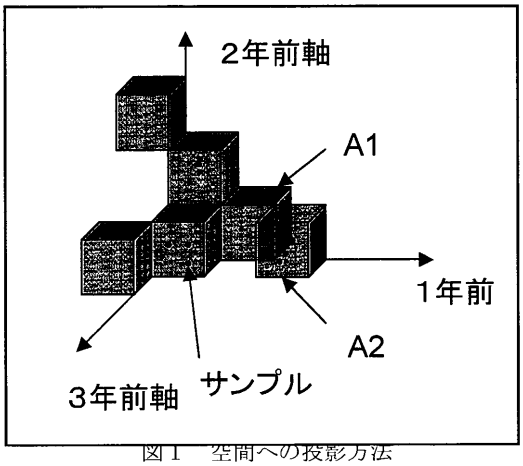

図１　空間への投影方法

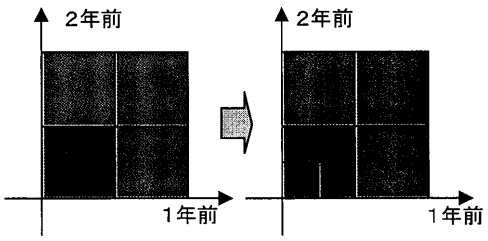

図２　分割領域の分割方法

③ それぞれの分割領域の中のサンプルの個数を数える。
④ Confidence：ラベルありの人数÷（ラベルありの人数＋ラベルなしの人数）
　Support：ラベルありの人数÷データの総数
で計算する。

4．利用したデータ

最低血圧：疾病ラベルなし　8081 件
　　疾病ラベルあり　152 件
最高血圧；疾病ラベルなし　8088 件
　　疾病ラベルあり　151 件

① それぞれの３分の２を教師データ、残りの３分の１をテストデータに利用する。
② 教師データに相関ルールを利用して confidence を求める。
③ テストデータを教師データで求めた空間に適用する

1．東京工業大学像情報工学研究施設
2．東京工業大学総合理工学研究科
3．東京工業大学フロンティア共同研究センター
4．バイオコミュニケーションズ株式会社

ことで②の妥当性を確かめる。

5．結果

≪最低血圧の検査値推移波形と高血圧症について≫
今回の発表では閾値：0.11(%)と設定したときに、閾値を満たす波形の中で confidence の高い波形の図と考察用の図を載せる。図は２つの線の間に収まるような最低血圧の検査値推移が起こった場合に、教師データの confidence の確率で翌年高血圧症に罹患することを示す。

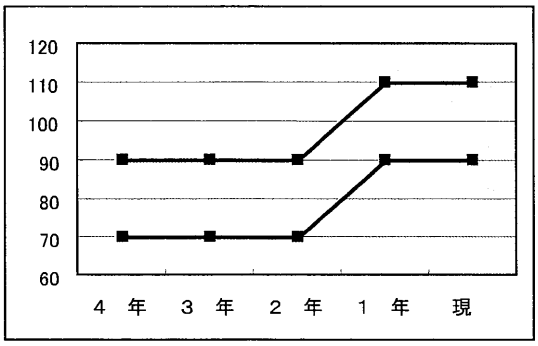

図３　confidence が１番目に高い波形
教師データ confidence：46.7(%)
教師データ support：0.13(%)
テストデータ confidence：75.0(%)
テストデータ support：0.11(%)

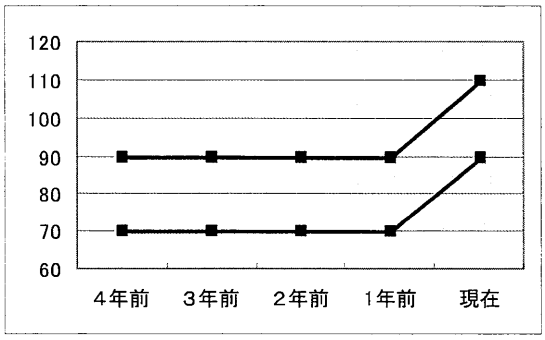

図４　考察のための波形
教師データ confidence：20.4(%)
教師データ support：0.20(%)
テストデータ confidence：30.6(%)
テストデータ support：0.40(%)

6．考察

- 今回の実験では図３の場合の波形が confidence 最大となった。２年前から１年前にかけて血圧値が上昇する検査値推移の波形である。
- 図４のように検査値が上昇したのち、その次の年も 90 から 110 の間の検査値である場合、教師データの confidence は 20.4(%)から 46.7(%)に上昇し 2.3 倍高血圧症に罹患する危険性がある。
- また図４と比較して、２年前から１年前にかけて１度検査値が 90 から 110 の間に上昇したのちに 70 と 90 の間に下降する場合の検査値推移の波形もある。（図は省略）そのときの confidence は約 7(%)となり高血圧症に罹患する危険性が大きく下がる。
- 上記のように検査値がいったん上昇してから下降する場合でも、全領域での confidence である 1.8(%)よりは 4 倍高血圧症に罹患する危険性が高いことも分かる。
- 現在までの研究では、過去よりも現在に近くなるに従って検査値が上昇する場合が高血圧症に罹患しやすいという当然の結果が相関ルールを利用することでも得られている。
- また、検査値が上昇する推移波形でも過去の検査値の値が低い場合 confidence は低くなることが分かる。
- 図３の場合のように教師データとテストデータの confidence が異なる場合があるが、これは教師データとテストデータに分ける際に、偏りが生じたためである。

7．まとめ

本研究では相関ルールを利用することで、血圧の検査値推移波形と高血圧症の関連性を調べた。相関ルールを利用するときに、support の値を利用して時系列の波形を投影した空間を分割する手法を利用して解析を行った。本研究の方法で時系列方向の波形を解析することで、検査値推移と生活習慣病との関連を理解できるのではないかと考えている。

今後の課題として、（ｉ）部分空間に分割する際の support の閾値の決定方法の検討（ii）テストデータを教師データに適用したときの検討方法（iii）BMI や問診結果などを取り入れた複数項目への検討などに取り組んでいきたいと考えている。

8．参考文献

[1] 山口雅浩,大倉恵子,大山永昭,大前和幸,舟谷文男,古海勝彦,小林祐一,佐々木敏雄「時系列健康診断データの確率密度分布に基づく疾病リスクの定量化」産業衛生学雑誌 42 臨　pp608(2000)
[2] Gautam Das,King-Ip Lin,Heikki Mannila,Gopal Rengannathan,Padhraic Smyth “Rule discovery from time series”
[3] Rakesh Agrawal,Tomasz Imielinski,Arun Swami “Mining Association Rules Between Sets of Items in Large Databases”