ID3を用いた受診行動データの特徴抽出とその視覚化ツール

Feature extraction and visualization tool from consultation behavior data using ID3

前田太陽[1)]，十倉伸太郎[2)]，松原光也[3)]，村田忠彦[4)]
Taiyo Maeda, Shintaro Tokura, Mitsuya Matsubara and Tadahiko Murata

1) 博士（理学） 関西大学 政策グリッドコンピューティング実験センター
（〒564-8680大阪府吹田市山手町3-3-35，maeda@pglab.kansai-u.ac.jp）
2) 関西大学大学院 総合情報学研究科 知識情報学専攻
（〒569-1095大阪府高槻市霊仙寺町2-1-1，stokura@pglab.kansai-u.ac.jp）
3) 修士（文学）関西大学大学院 文学研究科 地理学専攻（〒564-8680大阪府吹田市山手町3-3-35）
4) 博士（工学） 関西大学 ソシオネットワーク戦略研究機構（〒564-8680大阪府吹田市山手町3-3-35）

In order to recognize patient behavior for children in community healthcare, we develop a tool visualize medical treatment activities. Those activities are based on survey results we have conducted in Suita city, Osaka, Japan. We visualize them after extracting features of data sets using ID3. As a result, we obtain some classification rules. We develop a visualization tool on a GIS system in order to show patient behaviors on a map.

Key Words : ID3, visualization, medical treatment activities

1. 概要

近年の地域医療において，医療従事者の過重労働などいくつかの問題が存在する．地域医療の充実には，地域住民，医療従事者，行政による取り組みが必要であり，問題の迅速な分析や医療資源の効率的な管理が求められる．これらは，現状認識や課題を明らかにする必要があり，医療提供の状況，また地域住民に対する調査データなどと地域情報を関連付けた検討や考察が望まれる．この実現手段の1つとしてGISによる空間分析が役に立つ．調査データなどのアンケート結果に対し，地域医療の改善のためのデータ分析は，大きく分けて，データ全体に着目する統計学的な観点からの分析と，地域別の特定のデータに着目する地理的な観点から分析する場合がある．本研究では，患者に対する受診行動の分布と，医療従事者の必要性や過重労働などの状態を把握するため，地域的な分析を行う特徴抽出を目指した．

空間分析において，多数のサンプルから迅速に特徴をつかむことが，分析の際の１つのニーズである．この際，患者側の視点に立った患者の傾向をつかむことが重要であり，稀有な受診行動を持つ患者がどのように分布しているかを発見することも同様に重要である．しかしながら，GIS上でサンプル数が増えた場合，属性を含む患者個別の受診行動の特徴が見づらいといった問題が存在する．

そこで本研究は，If-Thenルール生成法の1つとして用いられるIterative Dichotomiser 3 (以下，ID3)[1)] を用い，患者の行動ルールを生成し，ルールにより調査データをグループ化することで，部分的な視覚化を行った．本稿では，ルール生成において特徴抽出のためのデータの前処理やID3の適用方法を述べ，抽出されたデータから得られた特徴と空間分析への適用方法を示す．

2. 医療機関選択アンケート

本研究で取り扱った受診行動データは，小児医療や少子化対策に関わるアンケート調査[2, 3)]から得られた個票データとした．各サンプルは属性と属性値を持つ．その詳細の一部を表1に示す．この調査は，2006年6月に大阪府吹田市の6つの幼稚園を対象に行い，いずれかの幼稚園に属する各家庭に対して実施し，回答者数1492名，回答率68.9%であった．アンケートは選択型回答形式で行った．さらに3章で説明するID3で扱うデータとして，先に述べた属性に加えて，患者それぞれの郵便番号から，その町丁目の中心座標を得て，その中心を患者の住所とみなし，その座標と選択した医療機関の距離を，自宅から医療機関までの直線距離として属性の1つとした．

次に，患者の行動特性としてどういったものが見たいかを考えると，地域や個人属性ごとにどの医療機関を選択している傾向が強いかがわかれば，患者の集中度合いがわかり，それらの傾向から，地域に関する議論への発展が考えられる．また，集中傾向と異なる患者がどれほどいるのかといった内容が分かれば，稀な受診であるように見えるが，どれほどの措置が必要かという議論が可能となる．

これらを実現する方法を考えた場合，着目したい属性項目に関して分析することもできるが，属性数と属性数が増えるにつれて膨大な組み合わせを考慮しなければならない．本稿では多変量分析の一手法を用い，調査データの細分化を簡略化することを議論の対象とする．なお，調査データの全回答結果は1492であったが，その中で表1の属性を全て持った回答は497であった．

表1.　アンケート調査結果の属性と属性値

交通手段
1.徒歩，2.自転車，3.自家用車，4.電車・バス
自宅からの通院時間
a.5分未満，b.5分～10分未満， c.10分～15分未満，d.15分～20分未満， e.20分～25分未満，f.25分～30分未満， g.30分以上
クチコミ（公共施設と民間施設それぞれ）
1.ある，2.なし
月収（妻と夫それぞれ）
a.収入なし，b.5万円未満 c.5万～10万円未満，d.10万～20万円未満， e.20万～30万円未満，f.30万～40万円未満， g.40万～50万円未満，h.50万～60万円未満， i.60万～70万円未満，j.70万～80万円未満， k.80万円以上
年齢（妻と夫それぞれ）
a.20歳代，b.30歳代，c.40歳代，d.50歳代
幼稚園名
6つの幼稚園名 K1,...,K6
自宅の郵便番号
67区域の郵便番号 P1,...,P67

3. 受診行動特性の抽出

調査データ等から，知識を獲得する手法は多く提案されており，それらの中でも決定木を用いたルールの生成手法は，分析後の構造の理解や解釈が容易である．また，獲得情報量の期待値を最大値とする属性を選択することで，決定木を構築する決定木構築法の代表的な手法であるID3を本研究で扱った．ID3は，分類効率のよい入力属性を選択できる利点[4,5,6]を持つ．ここでは，決定木からIf-Thenルールの生成時に，条件部の属性をユーザによって決定し生成する方法と，ID3によって生成する方法について述べる．また，属性値のグルーピングによりルール数を削減させ，どのように調査データを分類するかについて述べる[7]．

3.1　決定木を用いた行動ルール

決定木によるIf-Thenルールは，最上位ノードから分岐と中間ノードを経て，最終ノードを結んだものが1つのルールとなる．ノードは属性を，分岐は属性値を表し，最終ノードにおける結論部は，患者の自宅から患者が選択した医療機関までの移動距離とした．これは地域医療政策や公平な医療サービスの提供の観点から，医療圏をキーワードとした地域医療に関する議論が多いためである．このことから，If-Thenルールの条件部に使用された属性に基づいた結論部は，自宅から医療機関までの距離区分とし，400m未満，400mから800m未満，800mから2km未満，2km以上の4クラスに分類した．結論部の決定は，そのノードに所属するサンプル数が最大の距離区分とした．最大サンプル数を与える距離区分が複数の場合は結論部を保留とした．

次に生成されたルールの識別評価をおこなうため，If-Thenルールの結論部と調査での回答が一致した場合を正答，異なる場合を誤答，保留の場合は回答を保留とし，アンケート結果数からそれぞれの割合を算出したものと，説明ルール数を基準とした．まず，調査アンケートにおいて各属性が全体の結果に対してどれほど影響しているかを把握するために，1属性のみでの正答率を計算した結果を図1に示す．図1から属性の中で最も正答率に影響する属性は郵便番号であることが確認できる．

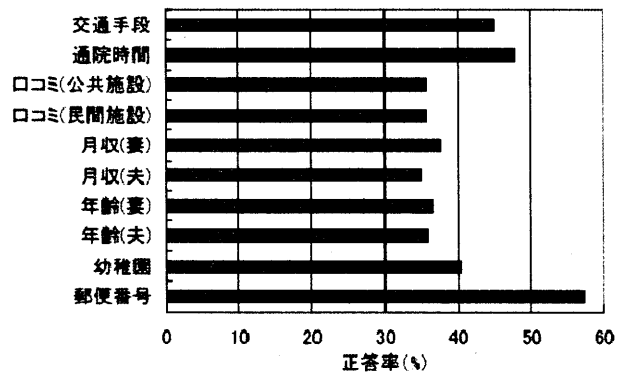

図1　調査アンケート結果全体に対する1属性の影響

3.2　ユーザ選択によるルール

ユーザが一意に属性を決定し，属性値から導きだしたルールのパターンは，属性と属性値に大きく左右され，そのパターン数は属性数の指数乗に増大する．そこで，単純に理解しやすいことを前提とし，ルール数を制限するため，決定木で扱う属性を最大で3階層とした．条件部に用いる属性の組み合わせは郵便番号を除く9属性を扱う場合，${}_9C_3=84$通りあり，全10属性での場合${}_{10}C_3 10=120$通りとなる．

属性の階層を3階層とした場合のルールの識別率を表2，表3を示す．表内の結果は，ユーザにより選択される3つの属性の組み合わせによる正答率の最大と最小のときの識別率である．アンケートの回答者によっての属性に対する回答の有無があるため，サンプル数が異なる．アンケートの回答の有無の検証のため，属性に対する全ての回答がある場合とない場合の比較と，郵便番号の属性を省いた9属性と郵便属性を含む10属性での結果を比較す

る．表2と表3において，正答率はほぼ同じであったが，ルール数の変化が大きく見られ，全ての属性を持ったアンケート結果は，少ないルール数での説明ができることが分かる．一方，郵便番号属性を入れることで正答率が良くなる半面，ルール数が多くなった．

表2 全ての属性を持たないアンケート結果を含むユーザ選択によるルールでの識別率

	9属性		10属性	
	最大	最小	最大	最小
サンプル数	761	683	740	694
正答率	56.8%	32.8%	81.8%	59.1%
誤答率	37.2%	60.0%	8.2%	24.1%
保留率	6.0%	7.2%	10.0%	16.8%
ルール数	99	13	528	181

表3 全ての属性を持つアンケート結果のユーザ選択によるルールでの識別率

	9属性		10属性	
	最大	最小	最大	最小
サンプル数	497	497	497	497
正答率	56.7%	31.6%	80.9%	59.8%
誤答率	36.2%	54.1%	5.2%	18.9%
保留率	7.1%	14.3%	13.9%	21.3%
ルール数	73	11	328	149

ここで問題となる点が2点ある．正答率に影響する郵便番号の属性をどのように扱うかと，ユーザが選んだ属性の組合せ全てから最適な決定木を選択しなければならない点である．

3.3 ID3による行動ルール

ユーザ選択によるルール生成での問題を解決するためID3を用いて決定木を構築する．表1の属性値を全て持つアンケート結果に対してID3を用いたルールでの識別率を表4に示す．表3と比べ属性抽出と決定木構築にID3を用いることで，ルール数を減少させることができた．次によりルール数を少なくし，説明を容易に理解できる形で，アンケート結果の分類をより緩めることを試みた．その方法は，異なる属性値をまとめることでそれらのルールが同一ルールになる性質から，次の2つの方法で行った．

3.3.1 主要7地域のグルーピング

大阪府吹田市は，鉄道や幹線道路を元に大きく7地域に分類することができる[8]．地域属性を考慮し，郵便番号情報をこの主要7区分でグルーピングした．この地域属性を用いID3により得られた結果を表5に示す．表4に対して，地域属性のグルーピングによりルール数を削減することができた．しかしながら，地域属性の影響を緩めたことで，正答率が下がり，地域の特徴が薄れたルールが存在することが考えられる．

3.3.2 属性値のグルーピング

3.3.1での地域属性のグルーピングでの問題を考慮し，属性のグルーピングを以下のように行った．図1でのIf-Thenルールにおいて，結論部が同じで，属性値が隣接する属性値を新たな属性値とした．対象とする属性は通院時間，月収，郵便番号である．この方法によるルールの識別率を表6に示す．これにより，正答率を大幅に向上させ，表3の10属性での最小ルール数の決定木よりもルール数を削減し，なおかつ正答率を向上させることができた．

表4 ID3によるルールでの識別率

サンプル数	497
正答率	66.4%
誤答率	17.3%
保留率	16.8%
ルール数	151

表5 主要7地域でグルーピングした識別率

サンプル数	488
正答率	55.7%
誤答率	40.0%
保留率	4.3%
ルール数	44

表6 属性値をグルーピングした識別率

サンプル数	488
正答率	70.4%
誤答率	22.7%
保留率	6.9%
ルール数	112

4. ID3によるルール生成と視覚化

4.1 ルール結果の視覚化

視覚化で扱うデータは，患者がどの医療機関を選択したかという受診行動，医療機関情報，患者の自宅の地域情報の3つであった．受診行動は，アンケートの回答者個別のID，居住区域の郵便番号，医療機関，個人属性を持つ．医療機関情報は，個別のID，医療機関名と地理的な座標である．地域情報は２つあり，道路情報と，町丁目地域情報[9]である．町丁目地域情報は，町丁目名，郵便番号，その地域の中心座標を持つ．地域別の動向を見るために，スパイダーダイアグラム[10]で表現するシステムを開発した[11]．このシステム上で，3.3節でのID3により分類した結果から，患者の居住地域の中心と選択した医療機関の2点を，半透明の線で表現した（図2, 図3）．円の中心が医療機関，円の半径が受診数である．複数の患者が同じ町丁目から特定の医療機関を選択している度合いを認識できる．

図2より，ルールに該当する患者の傾向については，ルールとその正答率から特徴が得られ，図3から，ルールに該当しない患者にも，そのルールには該当しないが特定の理由による医療機関の選択に傾向があることがつかめた．視覚化において，属性の組合せを条件部とし，どのようなデータ処理を行ったかによって，結論部のデータが一意に決まることから，それらの情報と生成されたルールが，母集団のデータセットに対するフィルタと解釈できる．

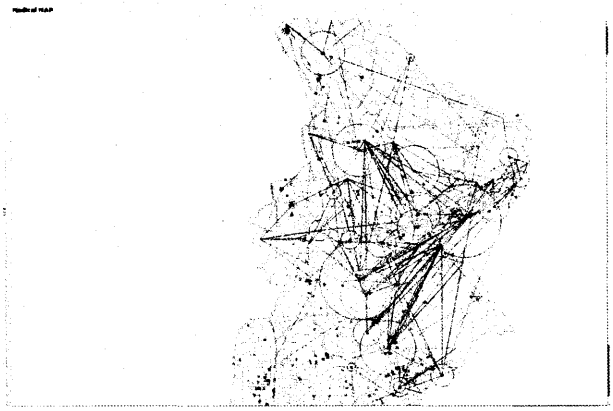

図2　ID3によるルールに該当する患者の行動の視覚化

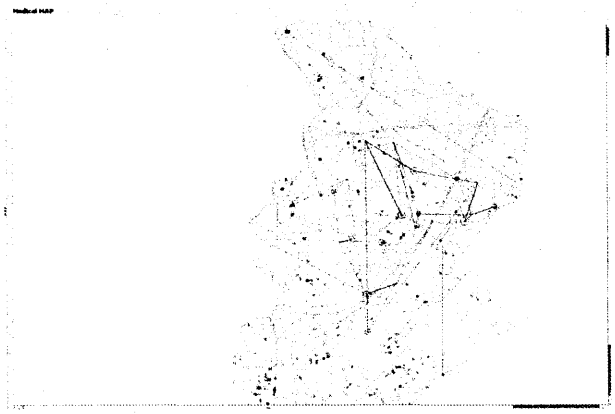

図3　ID3によるルールに該当しない患者の行動の視覚化

5．まとめ

本研究では，地域医療の状況把握するため，アンケート調査結果に基づき，ID3を用いてルール生成によってサンプルの分類を行い，視覚化を行った．

調査データのグルーピングにおいて，ユーザが選んだ属性を基にした行動ルール，ID3を用いた行動ルールによる分類方法を示し，ルールを特徴と見ることで特徴の抽出が可能になった．さらにそのルールに該当しないデータを見ることで，データのルールに属さない少数派の特徴を抽出することができた．生成されたルールにより誤答とされるサンプルを単に誤答として扱うのではなく，一般的なサンプルと異なる理由によって異なる行動をしているサンプルであると見なすことができる．本研究で開発したツールにより，ルールにより分類されたデータセットと分類されなかったデータセットを視覚化することで，特定の特徴を持ったグループの地理的分布を迅速に把握することができた．

今後は，ルールによって分類した属性と属性値と分類されたデータセットのデータベース化や視覚化の表現を強化したい．

謝辞：本研究の一部は，文部科学省社会連携研究推進事業（平成17年度～平成21年度）による私学助成を得て行われた．

参考文献

1) John Ross Quinlan, “Induction of decision trees, Machine Learning,” Vol.1, pp.81-106, 1986.
2) Shigeru Matsumoto, Yang Cao, "Resolving Service Quality Uncertainty through Word-of-Mouth Communication", Proceeding of International Conference of Socionetwork Strategies and Policy Grid Computing 2008, pp.95-108, 2008
3) 子育てアンケート調査報告書，関西大学 政策グリッドコンピューティング実験センター，2007
4) 馬野元秀: ID3，日本ファジィ学会誌，Vol.6，No.3，pp.502-504，1994.
5) 入月康晴，古橋武: ファジィエントロピーに基づくファジィID3の提案，日本ファジィ学会誌，Vol.14，No.3，pp.329-333，2002.
6) Jianbing Huo, Xizhao Wang, Mingzhu Lu, Junfen Chen, “Induction of Multi-stage decision tree”, Proc. of IEEE International Conference on Systems, Man, and Cybernetics, pp.835-839, October 8-11, 2006.
7) 十倉 伸太郎，村田 忠彦，医療機関選択アンケートからの意思決定ルールの抽出と同定，日本知能情報ファジィ学会 ファジィシステムシンポジウム 講演論文集，Vol.24，pp.58-62，2008
8) 松井幸一，水谷憲次，佐々木孝恵，松原光也，医療政策とGISによる可視化支援，PGLabディスカッションペーパーシリーズ，第10号，2006
9) 村田忠彦，鵜飼康東，政策グリッドコンピューティングとマルチエージェントシミュレーション，多賀出版，pp.101，2008
10) Naoko Nihei, Masahiro Yoshida, Hiroyuki Kaneta, Ryota Shimamura, Mutsuo Kobayashi, “Analysis on the Dispersal Pattern of Newly Introduced Latrodectus hasseltii (Araneae: Theridiadae) in Japan by Spider Diagram”, Journal of Medical Entomology, pp.269–276, 2003
11) 前田太陽，松原光也，村田忠彦，地域医療のためのデータ収集と視覚化ツールの開発，第11 回問題解決環境ワークショップ論文集，pp.43-46, 2008