データマイニング結果の可視化，比較・分析支援ツールの開発

三井田 浩† 　　和田 雄次‡‡
東京電機大学大学院 情報環境研究科†
東京電機大学 情報環境学部‡‡

1. はじめに

近年は情報社会とも呼ばれ，さまざまな種類の情報が溢れている．それに伴い大量のデータが蓄積されるようになった．今日では，そういった大量のデータを対象にルールや規則性を発見するデータマイニングを行う企業が増えている．企業が行うデータマイニングは，その企業が保有しているデータやその企業に関するデータが主な対象となる.データマイニングによって自動抽出されたルールは結果分析のために図や表に可視化され，それらから得た知見はビジネスに新たな利益を生み出すために活用される．しかし，分析者はマイニング結果が有効なものであるかどうかを自分で確かめる必要があり，それに慣れていない場合，どの部分が正しく，また有効なものかどうかを判断することが難しい[1].

そこで，本研究では，このマイニング結果の可視化方法に着目し，同一のデータをいくつかの可視化手法を用いて結果を出力し，それらを比較・分析する支援ツールの提案を行う．

2．マイニング結果と可視化

データマイニングの可視化方法として代表的なものに，決定木がある．これは，データマイニングを行うデータのある属性に関する重要な知識を，木構造によるルールの組み合わせで表現したものである．また，決定木は分類していくにつれ，枝やノードが増え，導き出されるルールは複雑なものとなる．決定木作成ツールではその点を考慮してか，全ての属性で分類はせずに割とシンプルな決定木が作成されることが多いが，それでも決定木には複数のルールが含まれている[2].

Making of data mining result and visible development of comparison and analysis supporting tool.
Hiroshi Miida, Yuji Wada
†Graduate School of Information Environment, Tokyo Denki University
‡‡Faculty of Information Environment, Tokyo Denki University

その中から有効なルールを発見するにはどうすればよいかが問題となる．

その解決方法として，同じデータから様々な可視化を行い，それらを比較するということが考えられる．また，可視化した結果からさらに別の分析を行うことで，マイニング結果の有効性が高まることが期待できる．

3 マイニング結果の可視化

まずは２つの方法でデータの可視化を行った[3]．データはWekaのSampleデータを用いた[4]. Sampleデータには属性，天候｛晴，曇，雨｝，風{有，無}，湿度｛real number｝，play{YES，NO}を含む．

１つ目の可視化方法はある属性に対して他の属性がどのような値を取っている場合が多いかを示したものである（図１）．ここではplay{YES，NO}の件数をそれぞれの属性の値ごとに何件あるかを表示している．図１では天候は曇，風は無，湿度は~70(%)と71~80(%)の時にplayがYESになる件数が多いことが分かる．現段階では天候，風，湿度のいずれか一つの属性を用い

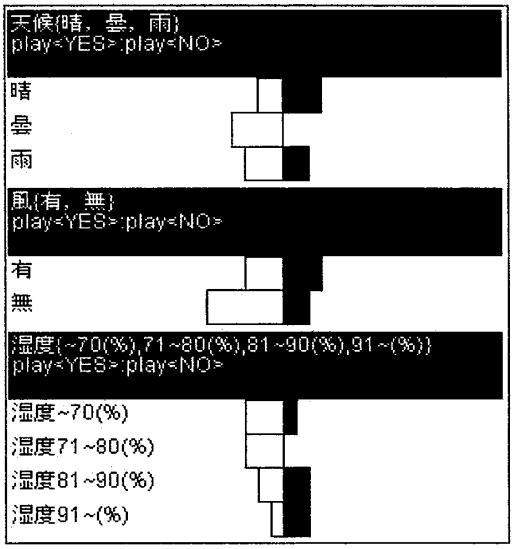

図１．複数の属性関係を示した可視化手法

て比較を行っている．

２つ目の可視化方法は属性ごとにその値が何件あるかを表示するものである（図２）．図２では天候，風，湿度のそれぞれの条件ごとの件数を表示している．

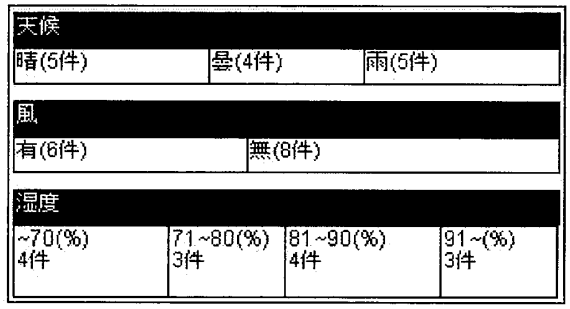

図２．属性値ごとのデータ件数

４．　マイニング結果比較・分析支援ツールの提案

分析者が決定木のルールが有効であるかどうか判断することが難しいことがある．その場合，可視化した結果から決定木に用いられている属性について比較すること，またその属性についてより詳細なデータを表示することにより，分析者の決定木の理解と有効なルールの認識につながる可能性が高まることが期待できる．ここでは可視化した結果が有効なルールであるかどうかを判断するマイニング結果比較・分析支援ツールを提案する．

４.１　属性選択機能について

ここでは分析者が比較対象となる属性の選択を行い，マイニング結果との比較を行う手法を提案する．

まず，分析者は属性ごとの全データの内訳を示した図１や図２から興味を持った属性値について調べる．図３は一例として分析者が{天候:晴}，{風:無}の２属性の属性値を選択した時，play{YES,NO}がどのようになるか調べた結果である．このように分析者が可視化された結果から興味を持った属性のデータについて，他の複数の属性との兼ね合いなど，その属性のデータにどのような傾向があるかを調べることによって，有効なルールであるかどうかを判断できる支援機能である．

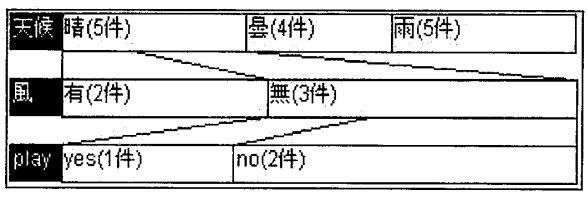

図３．属性間のつながり

４.２　　自動抽出機能について

ここでは4.1節で述べた属性選択機能で行う属性選択を自動で行う機能として２つの手法を提案する．

１つ目は分析者がある属性値を選んだ後に，比較対象を選択するのではなく分析者が興味を持った属性値とつながりの強い属性値を自動的に表示する手法である．これにより分析者の作業負担を軽減し，また意外な結果の発見につながることが期待できる．

２つ目は属性選択機能で比較対象にした属性の関連を調べる手法である．これにより分析者がそれまであまり注目していなかった属性のつながりを発見することができる．もしデータの属性数や件数が増えた場合，多数の分析対象が発生することが考えられる．そのような場合は,例えばその対象を分析したいマイニング結果(決定木など)に用いられている属性に限定することで対象を絞り込むことができる．また，今回は比較対象が件数のみであったが,これにサポート値やリフト値,確信度を追加することで新たな相関を示し,より分析者に理解しやすい結果になることが期待できる．

このように，分析者のマイニング結果比較・分析支援ツールには，マイニング結果と可視化したデータ，属性間のつながり,相関を用いる．

５．おわりに

今回は，WekaのSampleデータを用いて可視化を行い，可視化されたデータから有効な情報を発見するマイニング結果比較・分析支援ツールの提案を行った．今後は，このツールを使う際の一連の動作の流れをより詳細に定義し，必要に応じて新しい機能の追加を行う．また今後の課題として，今回用いたデータより件数・属性の多いもので比較・分析をできるようにしてデータに対する汎用性を持たせることや，可視化された結果を動的に表現することにより分析者がより結果を理解しやすい表示をさせることが挙げられる．

参考文献

[1]福田剛志，森本康彦，徳山豪，データマイニング，出版社(2001)

[2]石井義興，他．リアルタイム・データマイニングと相関関係の可視化，情報処理学会研究報告 Vol.2003 No.51(2003)

[3] How to use Weka tool box.
http://web.sfc.keio.ac.jp/~soh/dm03/man_w_03.html

[4]Ian H. DATA MINING, Practical Machine Learning Tools and Techniques, MORGAN KAUFMANN(2003).