D-030

データマイニング結果の比較・分析支援ツールの開発
Development of comparison and analysis supporting tool of data mining result

三井田 浩† 和田 雄次‡
Hiroshi Miida Yuji Wada

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，また分析者自身にとって，マイニング結果のどの部分が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，分析者がマイニング結果を理解し，より有効な選択を行えることを目的とした，元データとマイニング結果の比較・分析ツールの開発を行う．このツールは，可視化機能を含めた，元データとの関連を示す結果の比較・分析ツールである．また，可視化結果にユーザが直接操作を行うことができ，データの関連をわかりやすく行える機能を持つ.

2. マイニング結果の理解

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．一般的な決定木の例を図１に示す[2]．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．決定木には複数のルールが含まれている[3]．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，全ての属性は出力される決定木には含まれず，割とシンプルな決定木が作成されることが多い．しかし，図２のような決定木の場合，複数のルールを含む．このような決定木作成ツールにより作成された決定木の中から，分析者は有効なルールを発見し，それを有効活用するにはどうすればよいかが問題となる．

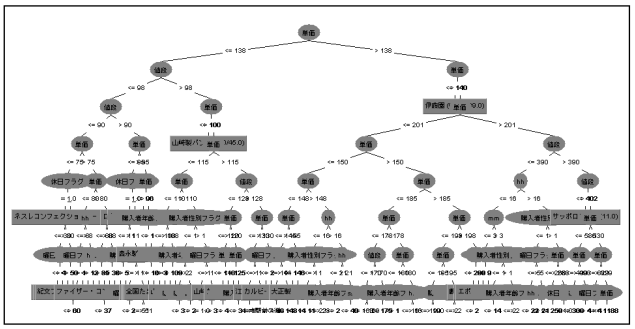

図１．決定木の例

3. データ比較

この章では，データマイニングやその元となったデータが持つ属性について説明をし，そのデータの扱い方に関する提案を行う．そして第４章以降で，今回開発を行ったデータマイニング結果の比較・分析支援ツールに関する機能説明，動作の流れといった具体的なツールの仕組みを述べる.

3.1 比較方法の提案

第２章で述べたように，データマイニング結果で導き出された結果は，分析者自身でその結果を理解し，有効なデータを選別し，そのルールを活用する必要がある.

その解決方法として，元データとマイニング結果で抽出された属性情報を比較する．それにより，元データに対してマイニング結果がどのような結果（例えば，他の属性との関連など），または違いなどを視覚的に判断することができる．それにより，マイニング結果の有効性が高まることが期待できる．また，分析者の操作により，分析者自身が興味を持った属性，相関といった情報から自動的に，ルールを作成し，提示する機能を提案する.

3.2 データ属性

属性を含むデータの例としてコンビニの販売履歴を図２に示す.

A	B	C	D	E	F	G	H	I	J	K	
立地	レシート番号	YYYY	MM	DD	曜日フラグ	休日フラグ	hh	mm	ss	購入者性別フラグ	購入
駅前	1	2003	9	1	1		4	39	8	2	
駅前	1	2003	9	1	1		4	39	8	2	
駅前	2	2003	9	1	1		7	48	58	1	
駅前	3	2003	9	1	1		8	47	10	2	
駅前	4	2003	9	1	1		8	47	57	2	

図２．コンビニの販売履歴

このデータには立地，取引された日時，購入商品など合計 20 の属性が含まれる．しかし，この中には決定木に含まれない属性もあり，分析者はその中から興味を引く属性などの情報があるかもしれない．また葉や枝の大きな決定

東京電機大学大学院 情報環境学研究科
情報環境工学専攻† 東京電機大学 情報環境学部‡
† Graduate School of Information Environment, Tokyo Denki University ‡ Graduate School of Engineering, Tokyo Denki University

木の場合，仮に属性が現れたとする．その属性が決定木から離れた別の属性との関連が知りたい場合，決定木だけでは，理解しにくい場合がある．

3.3 可視化機能

3.2 で述べた考えにより，元データから分析者が興味を持った属性を選択し，その属性に対する相関などの情報を可視化する．そして，その可視化情報にマイニング結果を合わせて表示し，比較することで，決定木に現れた相関と，自身で選択した属性間の相関の違いや特徴を理解することを目的とする．

そして，分析者がそのような操作を繰り返すことにより，その履歴により様々な相関を出力できる機能の提案を行う．

4. データ比較ツール

第３章で述べたデータ比較に関する提案を元に実際にツール開発を行った．この章では，具体的な機能，操作の流れを説明する．

4.1 概要

今回開発を行う，元データとマイニング結果の比較・分析ツール(以下，データ比較ツール)のインタフェースを図３に示す．

図３ データ比較ツールのインタフェース

データ比較ツールではデータマイニング結果とその元となったデータを入力として使用する．分析者は読み込むファイルを選択し，そのファイルのデータの属性を選択し可視化を行う．可視化された結果に，データマイニングにより出力された結果を合わせて可視化をする．現段階では，マイニング結果の読み込み方法としてはデータマイニングツール Weka により抽出された決定木を用いる[4]．また可視化された決定木はテキストで表されておりそれを用いる．可視化する順番として，1 読み込んだデータ，２データマイニング結果とする．

可視化された結果から，分析者の操作により詳細を見ることができる．この操作を繰り返すことにより，分析者のデータとマイニング結果を理解できることを目的とする．また，今回は開発に Java 言語を用いた．

4.2 機能と動作の流れ

4.2.1 データ読込

まずは可視化を行いたいデータを分析者が選択する．その画面を図４に示す．

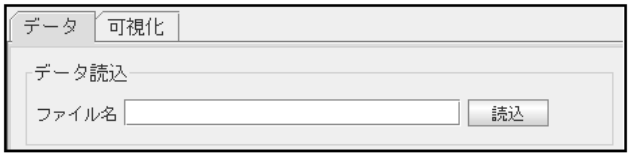

図４ データ読み込み画面

現在は，CSV ファイルとしてのみ選択可能であるが，今後はデータベースに保存されているデータも適用可能にする必要がある．図２のような CSV ファイルを読み取り，そのファイルの１行目を属性名として認識し，それを分析者がツールを用いる上で利用する属性名とする．この場合，属性名を見た時に何を表しているのか，すぐに理解できない属性名が含まれる場合がある．そのため，データの属性をある程度認識している分析者には，ツールで使用する属性名の変更ができる機能などの追加を，データを理解していない分析者には属性名を判別可能とするような機能を検討し対応が必要である．

4.2.2 データ選択

図３のデータ選択パネル内には属性選択とデータ選択の２つの項目がある．これらはデータ読込によって読み込まれたデータから，ユーザが興味を持った属性を選択し，可視化するデータを選択するものである．

ここで，図２のコンビニの販売履歴を持つファイルを読み込み，このファイルに含まれる属性がコンボボックスに追加された例を図５に示す．

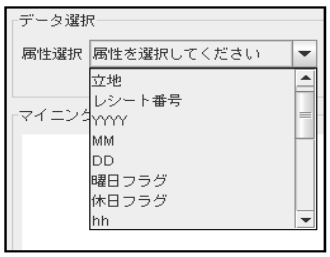

図５ 属性選択画面

図２のコンビニの販売履歴データの１行目にある属性，立地，レシート番号，YYYY，MM，DD，曜日フラグ，休日フラグなどが図５の属性選択画面内のコンボボックスに追加されていることがわかる．またスクロールにより図５には示されていないが，そのほかに，単価，個数，値段，メーカー名，分類名などの属性を持つ．ユーザはこれらの

CSV ファイルにより読み取られた属性名を用いる．コンボボックスに表示される文字は，CSV ファイルから読み取ったデータの属性名である．

次にデータ選択から出力するデータ項目を決める．データ選択パネル右側のデータ選択項目から選択をする．図６に示すようにデータ選択は確信度，サポート値，リフト値の３つの中から選択する．また，データ読込項目により，「Sales.csv」ファイルが読み込まれ，データ選択項目で属性選択として「値段」，データ選択として「確信度」が選択されている様子がわかる．

図６データ選択時画面

4.2.3 可視化

4.2.2 節で行ったファイルの読込，データ選択に対して可視化を行う．4.2.2 節でユーザによって属性選択「値段」，データ選択「確信度」が選択された例を示したので，その入力から可視化を行った様子を図７に示す．

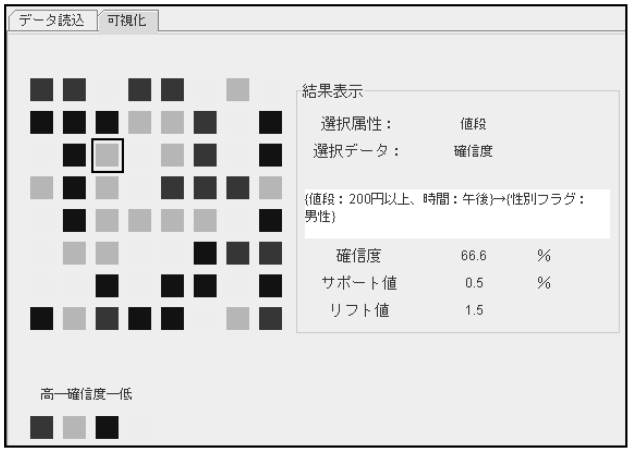

図７ 可視化とユーザによる操作例

図７の左側に行列でそれぞれ四角い図形が並んでいる．これはユーザが選択した属性「値段」を軸として２次元に様々なルールを並べ，確信度の高低により色分けをし，４種類に分類している．配置された属性の見方を図８に示す．導く相関ルールを(1)とする．

$$\{X,Y\}\rightarrow\{Z1,Z2,\ldots,Zn\} \quad (1)$$

ここで X は図８の階層構造のルート部分である「値段」となる．Y にはルートから１階層下の８つの属性が入る．このそれぞれ１属性が図７の可視化部分の１列となる．Z に関しては条件，1 つから複数の属性を含む．現在は確信度の最小値を予め指定し，その値を超えるものから順に表示しているが，可視化項目に含むかどうかを判定する，より詳細な決まりが必要となる．ここでの色分けは１列にそれぞれ８つのルールが表示されているが，その中の値の最大値と最小値を元に４分割をしている．

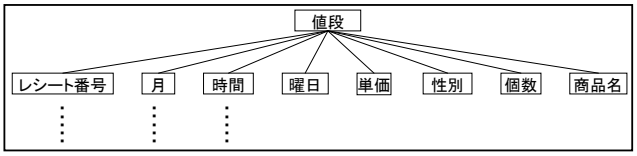

図８ 可視化された属性の階層構造

このように２次元配列に可視化されたデータに対し，ユーザは，それぞれの四角をクリックすることにより，詳細なデータを見ることができる．例として，ユーザが図７の左側部分の可視化結果から四角で囲まれた部分をクリックしたとする．すると，図７右側の結果表示パネル内にルールや相関を示す値など詳細なデータが表示される．クリックした場所は３列目なので，図８より値段と時間が(1)式の左部(条件部)となる．そして，結果表示パネルには，ユーザが選択した情報である「選択属性：値段」，「選択データ：確信度」が，その下にはユーザがクリックした位置のルールである「{値段：200 円以上，時間：午後}→{性別フラグ：男性}」が表示されている．これは，午後の時間帯に買い物の合計が 200 円以上であった客のうち，男性客の割合がどのくらいであるか，ということを示している．その結果がその下に表示されている確信度，サポート値，リフト値である[1].

確信度は上記の条件{値段：200 円以上，時間：午後}のうち，男性の割合を示すものである．ここでは 66.6%となっているので条件を満たす客のうち３人に２人が男性であることがわかる．次にサポート値であるが，これは全データの中で条件を満たす割合を示している．図７では 0.5%となっている．そのため対象の条件である{値段：200 円以上，時間：午後}→{性別フラグ：男性}は全データの 0.5%に当てはまるということである．最後にサポート値は，結論部分がどのくらい条件部と結びつきが強いかを示している．つまり，図７の結果では男性が来る確率よりも {値段：200 円以上，時間：午後}を満たす客のうち，男性である確率の方が 1.5 倍大きいということである．ただし，これは条件が起こる確率であり，実際の人数が 1.5 倍なわけではない[5].

マイニング結果を用いた比較について述べる．図３のインタフェースにマイニング結果を入力する様子を図９に示す．その結果，図７で行った可視化にマイニング結果を合わせて表示する．ここではデータマイニング結果として，データマイニングツール Weka を用いて Sales.csv に対しデータマイニングを行い，その結果を用いている[4].

図９マイニング結果入力

図９に示すように，マイニング結果を入力し，再度可視化ボタンを押すと，図７にマイニング結果によって得られたルールが含まれる場合，その項目が強調表示される．その様子を図 10 に示す．

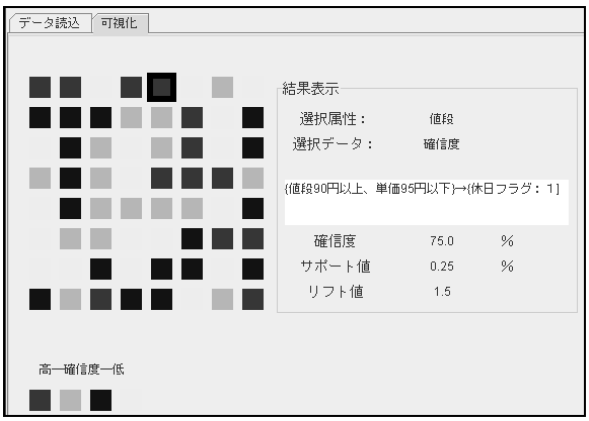

図 10 マイニング結果を合わせた可視化

１行５列目が他の四角よりも大きく強調されているのがわかる．これはマイニング結果として得られた相関ルールであることを示す．この可視化された情報に関するデータも図７で示した操作方法と同じようにクリックをすることで見ることができる．結果表示方式も同様である．

5. 追加機能提案

今回の目的であるユーザがデータマイニング結果から有効な情報(ルール)を選択することを支援するツールの機能をより具体化するために，追加機能の提案を３点ほど挙げる．

まず１点目に，現在は二次元に可視化をした後は，すべてユーザの判断で操作をしている．その場合，表示されるルールが多く，ユーザがなかなか目的の情報を得られないことが考えられる．そこで，可視化条件を指定する機能を追加する．例えば，相関を表す確信度の最小値を決める，あるいは，興味のない属性を選択し，可視化の対象から外すことなどが挙げられる．それにより，ユーザの操作時間が短縮するのではないかと考えられる．

２点目はユーザの操作履歴から，興味のある属性やその条件を判別する機能が考えられる．この場合ユーザが気になった情報を入力（例：５段階評価や興味がある or なしなど重み付けをする）し，それを元に相関を自動で求め，値の大きいものを表示する．

３点目は一般的に知られていないルールに的を絞り可視化を行う．用いるデータの中で一般的に明らかに既知が成り立つようなもの(例えば，ビールを買う人はおつまみを買う人が多いなど)をあらかじめ，グループとして分け，それらが相関ルールの条件部と結論部に含まれる場合は除外をする．ただし，相関ルールには基本的に複数の条件を含むので，グループ分けを行う定義が問題となる．

以上のような機能をデータマイニング結果から有効な情報を発見することを支援するために有効であるかどうかを検討していくことが今後の課題となる．

6. まとめ

今回はデータマイニング結果とその元データを，ユーザの操作によって可視化しながら比較を行う手法の提案とそのツールである元データとマイニング結果の比較・分析ツールの開発を行った．実際にデータを読み取り，可視化を行い，データマイニング結果を合わせて表示をする事が出来た．

問題点として，可視化結果が分かりにくい，クリックするまで内容がわからない．操作が終了するまでに時間がかかるのではないかといった点がある．今後の改良点として５章で述べたような機能の検討に加えて，インタフェースの改良，可視化結果がすぐにわかるような表示方法や機能を追加し，改良を行う．

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] How to use Weka
http://web.sfc.keio.ac.jp/~soh/dm03/man_w_03.html
[3] 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)
[4] Ian H. DATA MINING, Practical Machine Learning Tools and Techniques, MORGAN KAUFMANN(2003).
[5] JERICHO CONSULTING
http://www.jericho-group.co.jp/dic_dbm/kana/ri.html