Natto
情報理論的アプローチによる探索的データ解析ツール

Natto
A Tool for Exploratory Data Analysis with Information Theoretic Approach

鈴木 了太
Ryota Suzuki
株式会社 ef-prime
Ef-prime, Inc.
suzuki@ef-prime.com, http://www.ef-prime.com/

永井 達大
Tatsuhiro Nagai
（同　　上）
nagai@ef-prime.com

谷口 智也
Tomoya Taniguchi
株式会社 ef-prime，有限会社キメラワークス，東京工業大学
Ef-prime, Inc., Chimeraworks, Inc., Tokyo Institute of Technology
taniguchi@ef-prime.com, taniguchi@chimeraworks.jp, http://chimeraworks.jp/

keywords: exploratory data analysis, data visualization, knowledge discovery, association rules, entropy

Summary

Natto is an interactive data analysis tool for exploring associations in multivariate data. It provides graphical representation of data which assists intuitive understandings, and analytical tools such as cross-tabulation tables and association rule viewers. The strengths of associations between variables are measured by an index based on information theory, and represented by a directed graph. It can handle both quantitative and qualitative variables, and naturally handle missing values. In addition, an analyst does not need to specify any structure or assumption behind data. With these preferable characteristics, Natto provides wide range of users with easy way of data analysis. Moreover, it can play important roles in early stages of data analysis even by well-trained data analysts.

In this paper we first discuss an index called the *uncertainty coefficient*, which measures the strength of association between two variables based on information theory. We then discuss some indices to evaluate association rules including probability-based and entropy-based ones. The connection between an entropy-based measure called the *J-measure* and probability-based measures is discussed, as well as the relation between these measures and uncertainty coefficient. Next, our software named Natto is introduced. Natto is an exploratory data analysis tool which utilizes uncertainty coefficient, as well as signed J-measure and other measures. The usage of Natto is demonstrated using Fisher 's classical Iris data, and compared with classical approaches of multivariate analysis. Finally the advantages of our method are presented, and some tasks are stated as future works.

1. は じ め に

伝統的な統計学の問題設定は「よく計画された実験・観察により得られたデータからの仮説検証」にあり，確率分布論に基づいた厳密な仮説検定が展開された．これに対し，近年の探索的・発見的な問題設定においては「(しばしば分析とは無関係に) あらかじめ収集されたデータからの仮説探索」が問題とされることが多い．この新しい問題設定に対し，計算機科学の分野からは apriori [Agrawal and Srikant 1994] など種々の仮説探索アルゴリズムが生まれ，総称してデータマイニング手法と呼ぶようになった．この流れには統計科学の成果も大いに活用され，仮説探索における検定統計量の利用や情報量基準に基づくモデル選択，交差検証法による予測精度の評価などが応用されるようになった．

これらの方法論はソフトウェアパッケージに実装され，ゲノム科学や自然言語解析，マーケティングなど多くの分野へと応用範囲を広げていった．実際，問題がある程度明確であるような状況，すなわち予測対象となる目的変数やクラスタリングに利用するべき変数がわかっているような場合においては，データマイニングの方法論は効果的に働くことが多い．ところが，応用上の多くの問題においては「そもそもデータから何が判明し得るのか」「どのような観点から分析を行えば，データを有効に活用できるのか」といったことが事前にわかっていないような状況がしばしば起こる．したがって，データセット全体を俯瞰的に眺め，応用可能な何らかの関係性が潜んでいるかどうかを探索するための方法が求められる．

我々はこのような状況に対する解決策として「データを全体として俯瞰」し，かつ「局所的な関係性を対話的に抽出」するためのデータ解析ソフトウェア「Natto」を開発した．Natto は多変量データにおける変量間の相関関係を情報理論に基づく尺度で測り，有向グラフとして表現する．グラフはユーザーによってインタラクティブに操作可能であり，詳細な関係をクロス集計表や相関ルール探索によって確認することができる．ソフトウェアは無償で利用可能であり，我々のウェブサイト (`http://www.ef-prime.com/natto/`) からダウンロードできる．

本稿では，変数間の相関関係を表す指標について議論し，ソフトウェア Natto を用いた実際のデータ分析における応用例を紹介する．また，既存の方法論との比較を行い，今後の発展の可能性について議論する．

2. エントロピー基準に基づく相関尺度

2·1 問 題 意 識

以下，$n \times p$ 行列として表される表形式のデータ $\mathbf{D} = (d_{ij})$ について議論を行う．データ $\mathbf{D}$ を生み出す確率構造として p 次元の確率変数ベクトル $D = (D_1, D_2, \ldots, D_p)$ を考える．このとき D は同時分布 F に従うと仮定し，これを $D \sim F$ と表す．

ここで $\{D_1, D_2, \ldots, D_p\}$ から任意の 2 変数を取り出し，X，Y と表記する．このとき，X と Y の関連性の強さを同時分布 F_{xy} に基づいて測ることを考える．X，Y がともに量的変数である場合，相関係数 $\mathrm{Cor}(X,Y)$ がよく用いられる．相関係数は変量間の線形関係を測る指標であり，これが高い場合には強い相関関係が観測されるが，逆に強い相関関係があっても非線形な関係である場合には関係を検出できるとは限らない．また，データマイニングの応用場面においては質的変数と量的変数が混在する場合が多く，適用範囲が限られてしまうという欠点がある．

X, Y がともに質的変数である場合について理論的な考察を行った結果，我々は *uncertainty coefficient* [Theil 1970, Goodman and Kruskal 1979] と呼ばれる量に辿りついた．X の Y に対する uncertainty coefficient とは，X によって説明され得る Y の不確定性の程度をエントロピー基準によって測ったものである．0 から 1 までの値をとり，0 のとき X と Y は独立，1 のとき Y は X によって完全に説明される．概念的には単回帰分析における決定係数 R^2 と共通しており，同様の観点から利用することができる．また，変数間の局所的な関係を記述した「相関ルール（association rule）」についての評価指標である j-measure, J-measure [Smyth and Goodman 1992] とも関係があり，変数間の大域的な関係と局所的な関係を同じフレームワークで扱うことができるという利点がある．さらに，理論的には量的変数への拡張も可能であり，きわめて汎用性の高い指標といえる．本節ではこれらの指標について紹介し，相互の関連性や他の指標との関係について述べる．

2·2 変数間における相関関係の評価

ふたつの確率変数 X, Y の関連について，X の値を知ることによって減少する Y の不確実性の程度，すなわち X による Y の説明率を測る．簡単のため X, Y はともに離散確率変数とし，それぞれのとる値を $X \in \{x_1, x_2, \ldots, x_r\}$, $Y \in \{y_1, y_2, \ldots, y_c\}$ とする．確率ベクトル (X,Y) は同時確率分布 F_{xy} をもち，$P(X = x_i, Y = y_j) = p_{ij}$ と表記する．また，X および Y の周辺分布を $P(X = x_i) = \sum_{j=1}^{r} p_{ij} = p_{i.}$, $P(Y = y_j) = \sum_{i=1}^{c} p_{ij} = p_{.j}$ とする．

まず，X について何の情報も得ていない状態での Y の不確実性の程度を測る．これをシャノンのエントロピー [Shannon 1948] によって測ることにすれば，

$$\begin{aligned} H(Y) &= -\sum_{j=1}^{c} P(Y = y_j) \log P(Y = y_j) \\ &= -\sum_{j=1}^{c} p_{.j} \log p_{.j} \end{aligned}$$

となる．(ただし，$p_{.j} = 0$ のときは $p_{.j} \log p_{.j} = 0$ とする．以下同様) 次に，X が特定の値 x_i をとるとわかっているときの Y の不確実性，すなわち X を知った後に残った Y の不確実性を評価しよう．これは条件付分布 $P(Y|X = x_i)$ のエントロピーによって測ることができ，

$$\begin{aligned} H(Y|X = x_i) &= -\sum_{j=1}^{c} P(Y = y_j|X = x_i) \log P(Y = y_j|X = x_i) \\ &= -\sum_{j=1}^{c} \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \log \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \\ &= -\sum_{j=1}^{c} \frac{p_{ij}}{p_{i.}} \log \frac{p_{ij}}{p_{i.}} \end{aligned}$$

となる．実際には X はさまざまな値をとるから，X を知った後に残る平均的な Y の不確実性は，

$$\begin{aligned} H(Y|X) &= E_x[H(Y|X = x_i)] \\ &= \sum_{i=1}^{r} P(X = x_i) H(Y|X = x_i) \\ &= -\sum_{i=1}^{r} \sum_{j=1}^{c} P(X = x_i, Y = y_j) \log P(Y = y_j|X = x_i) \\ &= -\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \log \frac{p_{ij}}{p_{i.}} \end{aligned}$$

となる．Y の不確実性が $H(Y)$，X を知った後に残る Y の不確実性が $H(Y|X)$ であるから，X によって説明できる Y の不確実性の量を次のように定義できる．

$$I(Y|X) = H(Y) - H(Y|X)$$

$$
\begin{aligned}
&= -\sum_{j=1}^{c} p_{\cdot j} \log p_{\cdot j} + \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log \frac{p_{ij}}{p_{i\cdot}} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log \frac{1}{p_{\cdot j}} + \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log \frac{p_{ij}}{p_{i\cdot}} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}
\end{aligned}
$$

これは X と Y の相互情報量と呼ばれ，X と Y が共通にもつ不確実性の程度といえる．ここで $I(Y|X) = I(X|Y)$ であり，説明・披説明の関係を逆転しても値は変わらない．また $0 \leq I(Y|X) \leq H(Y)$ である．$I(Y|X) = 0$ となるのはすべての i,j について $p_{ij} = p_{i\cdot} p_{\cdot j}$ となるときであり，これは X と Y が独立であることを示す．逆に $I(Y|X) = H(Y)$ となるのは $H(Y|X) = 0$ のときであり，これは Y の X による条件付分布のエントロピーがゼロ，すなわち X の値が定まったときに Y の値が一意に定まることを示す．これは変数間の関連性の強さを測る指標として優れた性質といえる．理論的には $P(X,Y)$ と $P(X)P(Y)$ の Kullback-Leibler 情報量 [Kullback 1959] と一致し，「X と Y が独立である」という仮定のもとでのモデル $P(X)P(Y)$ が実際の同時分布 $P(X,Y)$ とどれだけ近いかを示す値として解釈できる．

相互情報量 $I(Y|X)$ は上記のような優れた性質を持つが，値の大きさが Y の不確実性の大きさである $H(Y)$ に依存してしまう．つまり，仮に確率変数 Z に対して $I(Z|X) = I(Y|X)$ であったとしても，そもそもの不確実性の大きさである $H(Y)$ と $H(Z)$ の大きさが異なれば，X による相対的な寄与の大きさは異なると考えるのが自然である．

したがって，相互情報量 $I(Y|X)$ を Y 自体の情報量 $H(Y)$ で基準化することを考える．すなわち

$$
\begin{aligned}
U(X \to Y) &= \frac{I(Y|X)}{H(Y)} \\
&= \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log \left[p_{ij} / p_{i\cdot} p_{\cdot j} \right]}{-\sum_{j=1}^{c} p_{\cdot j} \log p_{\cdot j}}
\end{aligned}
$$

とする．このとき $0 \leq U(X \to Y) \leq 1$ であり，X と Y が独立のとき $U(X \to Y) = 0$，Y が X によって完全に説明できるとき $U(X \to Y) = 1$ となる．定義から $U(X \to Y)$ は X によって説明できる Y の不確実性の割合として解釈することができ，$U(X \to Z) \geq U(X \to Y)$ であれば Y に比べて Z のほうが X によって説明できる不確実性の割合が高いということができる．

さて，実際に観測できるのは確率分布 F_{xy} ではなくデータ $\mathbf{D}$ であるから，データから $U(X \to Y)$ を推定する必要がある．ここでは，経験分布関数 $\hat{F}_{xy}$ によるプラグイン推定量を利用する．すなわち標本サイズを n，そのうち $X = x_i$ かつ $Y = y_j$ となった観測の数を $\#(X = x_i, Y = y_j) = n_{ij}$ とし，$P(X = x_i, Y = y_j) = p_{ij}$ を比率 $\pi_{ij} = n_{ij}/n$ で置き換えた

$$
\hat{U}(Y|X) = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} \pi_{ij} \log \left[\pi_{ij} / \pi_{i\cdot} \pi_{\cdot j} \right]}{-\sum_{j=1}^{c} \pi_{\cdot j} \log \pi_{\cdot j}}
$$

によって $U(X \to Y)$ の推定値とする．このとき $\hat{U}(X \to Y)$ は uncertainty coefficient [Theil 1970, Goodman and Kruskal 1979] と呼ばれる．

また，定義から $U(X \to Y)$ は連続確率変数に拡張することができる．確率関数の代わりに密度関数を用い，和を積分で置き換えればよい．しかし，一般の確率密度関数に対してエントロピーを計算することは容易ではない．分布が未知の場合には適切な確率モデルを構築し，データへの当てはめを行う必要があるうえに，積分の評価は解析的にも数値的にも困難な場合が多い．

したがって，Natto では連続変数をそのまま分析するのではなく，離散化することによって量的変数と質的変数の混在した分析を可能にしている．これについては後に詳しく述べる．

2·3 相関ルールの評価

変数間の局所的な関係を記述・評価するための方法として，相関ルール (association rule) [Agrawal and Srikant 1994] が用いられることがある．相関ルール は通常 $(A \to B)$ のような形式で表され，「事象 A が起こったという前提のもとで，事象 B も共に起こる」ことを表す．[Agrawal and Srikant 1994] における例を挙げると，スーパーマーケットにおいて「パンを買った顧客が，バターも同時に買っている」という場合，(パン → バター) のように表す．

これを確率変数間の関係として整理すると次のようになる．前節に引き続き，ふたつの確率変数を X，Y とする．事象 A，B はこれらの確率変数の値域における部分集合とすると，相関ルール $(A \to B)$ は $(X \in A \to Y \in B)$ と書き直すことができる（ただし，簡単のため以下では前者の記法を用いる）．これを先の例に当てはめてみよう．X は購入したパンの個数とし，$X, \in [0, \infty)$（ただし X は整数）とすると，$A = [1, \infty)$ と書ける．Y と B についても同様である．

§1 確率に基づいた指標

以上の準備のもとで，相関ルール $(A \to B)$ の評価について考える．相関ルール自体は単なる記述に過ぎない．応用上の観点からルールの「良さ」を定義し，良いと認められたルールをデータから抽出することを考える．[Agrawal and Srikant 1994] は相関ルールに関する 3 つの評価指標を定め，それらに基づいて高速にルール抽出を行う apriori アルゴリズムを提案した．[Borgelt and Kruse 2002] はこのアルゴリズムを高速に実装し，応用上の観点から指標の修正を提案した．以下にこれらの評価指標を紹介する．

先に述べたように，相関ルール $(A \to B)$ は「A が起こったとき，B も起こる」ことの記述であるから，A が

起こったときの B の生起確率が大きいほど良いルールであると考えられる．これは条件付確率 $P(Y \in B|X \in A)$ によって評価でき，**確信度（*confidence*）**と呼ばれる．前出の (パン → バター) の例では，確信度は「パンを買ったという条件のもとでバターを購入する確率」であり，これが高ければ「パンの近くにバターを配置すれば，一緒に売れる可能性が高い」といった推論を行うことができ，応用上の価値が高いルールといえる．

一方，パンを買った顧客の多くがバターも同時に購入するとしても，パンを買う顧客がほとんど存在しないのではルールの利用価値は乏しい．確信度はルールの絶対的な「強さ」を評価するが，これに加えてルールの「適用可能な範囲」を評価する必要がある．これを評価する指標が**支持度（*support*）**であり，A の生起確率 $P(X \in A)$ によって評価する[*1]．(パン → バター) の例のように，抽出されたルールをヒントとしてルール通りの事象を引き起こすことが目的とされるような場合，支持度は確信度と並んで重視すべき指標である．

ところで，確信度によってルールの「絶対的な」強さを測ることができるが，これには落とし穴がある．あるルールについて，確信度 $P(Y \in B|X \in A)$ が 80%であるとしよう．これは十分に高い数値に思えるが，$X \in A$ という条件を外した，そもそもの B の生起確率 $P(Y \in B)$ が 90%だとしたらどうだろうか．事象 A という条件を与えたことで，B の生起確率はかえって低くなってしまったことになる．これでは良いルールとはいえない．

そこで，ルールの強さを「相対的に」評価する指標として**リフト（*lift*）**が導入された．リフトは条件付確率と周辺確率の比，すなわち

$$\frac{P(Y \in B|X \in A)}{P(Y \in B)}$$

として定義される．「条件 A を加えることで事象 B が何倍起こりやすくなるか」と解釈することができ，リフトが 1 より大きいときルールは有効であると考えられる．また，リフトの定義を展開すると

$$\begin{aligned}\frac{P(Y \in B|X \in A)}{P(Y \in B)} &= \frac{P(X \in A,Y \in B)/P(X \in A)}{P(Y \in B)} \\ &= \frac{P(X \in A,Y \in B)}{P(X \in A)P(Y \in B)}\end{aligned}$$

となる．事象 A と B が独立に起こると仮定した場合，分母と分子は一致して 1 となることから，リフトは A と B の独立性を測る指標としても解釈できる．これらの指標を用いて，相関ルールの良さを総合的に評価することができる．

[*1] これは [Borgelt and Kruse 2002] による定義であり，オリジナルの [Agrawal and Srikant 1994] では $P(X \in A,Y \in B)$ を利用している．詳細な議論は apriori オンラインマニュアル `http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/apriori/apriori.html` を参照．

§2 エントロピーに基づいた指標

すでに説明した通り，確率に基づいた指標によって相関ルールを総合的に評価することができる．これらは応用上の必要性と強く結びついているうえに解釈が容易であるため，分析者のみならず分析結果の利用者にとっても利用価値が高い．しかし一方で，3 種類の指標を組み合わせて利用するため，複数ルール間の優劣をつけることが困難であるという欠点がある．

情報理論に基づいた単一のルール評価指標として，[Smyth and Goodman 1992] は J-measure を提案した．これは相関ルールの価値をエントロピーによって測ったものであり，形式的には確信度，支持度，リフトを総合した指標となっている．以下では J-measure の導出や他の指標との関連について議論し，前節で紹介した uncertainty coefficient との関係についても触れる．

では，相関ルール $(A \to B)$ をエントロピーに基づいて評価することを考える．評価の対象となるのは条件 $X \in A$ を与えたときの Y の不確実性であるが，ここでは $Y \in B$ であるか $Y \notin B$ であるかの二値的な問題として考える．これを形式的に $Y \in B$ のとき 1，$Y \notin B$ のとき 0 をとる確率変数 $\tilde{Y}$ を用いて記述しよう．このとき，$P(\tilde{Y})$ と $P(\tilde{Y}|X \in A)$ の「近さ」を Kullback-Leibler 情報量によって測ったものを j-measure といい，次のように定義される．

$$\begin{aligned}j(A \to B) = {} & P(Y \in B|X \in A)\log\frac{P(Y \in B|X \in A)}{P(Y \in B)} \\ & + P(Y \notin B|X \in A)\log\frac{P(Y \notin B|X \in A)}{P(Y \notin B)}\end{aligned}$$

J-measure はこれに $P(X \in A)$ を掛けたものとして定義される．

$$\begin{aligned}J(A \to B) &= P(X \in A) \times j(\tilde{Y}|X \in A) \\ &= P(X \in A,Y \in B)\log\frac{P(Y \in B|X \in A)}{P(Y \in B)} \\ &\quad + P(X \in A,Y \notin B)\log\frac{P(Y \notin B|X \in A)}{P(Y \notin B)}\end{aligned}$$

次に，J-measure と前出の確率に基づいた指標との関係について述べる．$J(A \to B)$ の第一項を展開すると，

$$\begin{aligned}&P(X \in A,Y \in B)\log\frac{P(Y \in B|X \in A)}{P(Y \in B)} \\ &= P(Y \in B|X \in A)P(X \in A)\log\frac{P(Y \in B|X \in A)}{P(Y \in B)} \\ &= (\text{確信度}) \times (\text{支持度}) \times \log(\text{リフト})\end{aligned}$$

のように展開できる．第二項では逆のルール $(A \to B^C)$ についても同様の評価を行っているため直観的な解釈は難しいが，確信度，支持度，リフトを総合した形式の指標となっていることがわかる．

前節で紹介した uncertainty coefficient と J-measure の関係についても触れておこう．$\tilde{Y}$ と同様，$X \in A$ のとき

1，$X \notin A$ のとき 0 をとる確率変数 $\tilde{X}$ を導入すると，$\tilde{X}$ と $\tilde{Y}$ の相互情報量は $I(\tilde{Y}|\tilde{X}) = J(A \to B) + J(A^C \to B^C)$ であり，したがって

$$U(\tilde{X} \to \tilde{Y}) = \frac{J(A \to B) + J(A^C \to B^C)}{H(\tilde{Y})}$$

となることがわかる．このことから，uncertainty coefficient が大きいということは変数間に強い相関ルールが存在することと関連づけて解釈することができる．

また，J-measure は事象間の独立性に関する指標であり，効果の正負は評価されない．つまり，定義より $J(A \to B) = J(A \to B^C)$ であるため，ルール $(A \to B)$ が起こりやすいのか逆に起こりにくいのかを判定することができないのである．Natto ではこれを解決するため，符号を返す関数 $\mathrm{sign}(\cdot)$ を用いて

$$J_{\pm}(A \to B) = J(A \to B) \times \mathrm{sign}\left(\log \frac{P(Y \in B|X \in A)}{P(Y \in B)}\right)$$

として定義した符号付 J-measure を利用している．

本節の最後に，X，Y がともに離散確率変数であるときの相関ルール $(X = x_i \to Y = y_j)$ の評価について整理しよう．記法については前出のものを用いる．

まず，確信度は

$$\begin{aligned}\mathrm{Conf}(X = x_i \to Y = y_j) &= P(Y = y_j|X = x_i) \\ &= \frac{p_{ij}}{p_{i.}}\end{aligned}$$

と書ける．実際にはデータ $\mathbf{D}$ から推定して，

$$\begin{aligned}\widehat{\mathrm{Conf}}(X = x_i \to Y = y_j) &= \hat{P}(Y = y_j|X = x_i) \\ &= \frac{\pi_{ij}}{\pi_{i.}}\end{aligned}$$

となる．以下同様に，支持度は

$$\widehat{\mathrm{Supp}}(X = x_i \to Y = y_j) = \hat{P}(X = x_i) = \pi_{i.}$$

リフトは

$$\begin{aligned}\widehat{\mathrm{Lift}}(X = x_i \to Y = y_j) &= \frac{\hat{P}(Y = y_j, X = x_i)}{\hat{P}(X = x_i)\hat{P}(Y = y_j)} \\ &= \frac{\pi_{ij}}{\pi_{i.}\pi_{.j}}\end{aligned}$$

となる．また，J-measure は，

$$\begin{aligned}&\hat{J}(X = x_i \to Y = y_j) \\ &= \hat{P}(X = x_i, Y = y_j) \log \frac{\hat{P}(X = x_i, Y = y_j)}{\hat{P}(X = x_i)\hat{P}(Y = y_j)} \\ &+ \hat{P}(X = x_i, Y \neq y_j) \log \frac{\hat{P}(X = x_i, Y \neq y_j)}{\hat{P}(X = x_i)\hat{P}(Y \neq y_j)} \\ &= \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i.}\pi_{.j}} + \sum_{k \notin j} \pi_{ik} \log \frac{\sum_{k \notin j} \pi_{ik}}{\pi_{i.}(1 - \pi_{.j})}\end{aligned}$$

であり，符号付 J-measure は

$$\begin{aligned}&\hat{J}_{\pm}(X = x_i \to Y = y_j) \\ &= \left(\pi_{ij} \log \frac{\pi_{ij}}{\pi_{i.}\pi_{.j}} + \sum_{k \notin j} \pi_{ik} \log \frac{\sum_{k \notin j} \pi_{ik}}{\pi_{i.}(1 - \pi_{.j})}\right) \\ &\times \mathrm{sign}\left(\log \frac{\pi_{ij}}{\pi_{i.}\pi_{.j}}\right)\end{aligned}$$

と書ける．

3. 探索的データ解析ソフトウェア Natto

この節では，本稿の主題であるデータ解析ソフトウェア Natto について述べる．分析例として Fisher の Iris データ [Fisher 1936] を用い，実際の利用方法を解説する．

3·1 特　　徴

Natto は「データを全体として俯瞰」し，「局所的な関係性を対話的に抽出」するためのデータ解析ソフトウェアである．多変量データにおける変数同士の関係を探索的に分析することを目的としており，GUI によるマウス操作を中心としたインタラクティブな操作を特徴とする．グラフの表示にはオープンソースの Java ライブラリ JUNG [O'Madadhain, et. al. (preprint)] を用い，高度なグラフ操作を実現した．

Natto の主な特徴は次の通りである．

(1) データセット全体を，変数をノードとし，相関関係を矢印で表現した有向グラフによって視覚化する．
(2) グラフはユーザーによってインタラクティブに操作することができる．例えば，ノードの位置を変更したり，矢印の表示に用いるパラメータを動的に変更することができる．
(3) グラフによって見えてきた変数間の相関関係を，局所的に確認することができる．グラフと連動したクロス集計表や相関ルール探索を利用することができ，相関ルール探索ではルールのグラフによる視覚化がなされる．相関ルールについても変数間の関係と同様にグラフによって視覚化され，インタラクティブな操作が可能である．

Natto はほとんどの操作をマウスのみで完結できるインターフェースを持ち，ユーザーに対して前提知識をほとんど要求しない．ソフトウェア上で利用している各種の指標は前節で展開した情報理論的アプローチに基づくものであるが，その解釈は「説明される不確実性の割合（説明力スコア；uncertainty coefficient）」「値が大きいほどインパクトが大きい（インパクト；符号付 J-measure）」のように極めてシンプルである．

上記の特徴によって，Natto はデータ分析の専門家のみならず，広く一般の応用分野において利用されることが期待できる．一般の統計・データマイニングソフトウェ

アのように関数関係の推定や統計的仮説検定を行うことはできないが，データ解析の初期段階においてモデリングのための「アタリ」をつけたり，調査・実験によって得られたデータから大まかに全体像を把握するといった場合において効果を発揮する．利用法の習得も容易であることから，従来データ解析に携わることが少なかった応用分野の担当者にも利用することができ，分析に対する現場における知識のフィードバックを促進することができる．

経験豊富な分析技術者にとっては分析の補助的ツールとして，応用分野の現場担当者にとっては従来に比べて手軽な分析ツールとして，データからの知識発見に活用されることを大いに期待する．

3·2 データの扱い

以下，Natto におけるデータの扱いについて述べる．Natto では全てのデータをカテゴリカルな変数として扱う．連続変数に対しては分割によるカテゴリ化を行い，四分位点による分割や応用上の意味に基づいた分割などをユーザーが自由に設定することができる． 図 1 は Fisher

	アイテム名	値	合計	パーセント
☑	Petal_width=1~3	1~3	41	27%
☑	Petal_width=4~13	4~13	38	25%
☑	Petal_width=14~18	14~18	38	25%
☑	Petal_width=19~25	19~25	33	22%

図 1 四分位点による連続変数の分割

の Iris データによる例である．`Petal_width` はアヤメにおける花弁の幅を表す連続変数であるが，ここではカテゴリの大きさがほぼ等しくなるように 4 分割されている．等間隔ではなく等サイズで分割することの理由としては，カテゴリのサイズが極端に小さくなることによる指標のバラツキを抑えることや，外れ値の影響を抑えてロバストな結果を得ることなどが挙げられる．したがって標準では連続変数は等サイズに分割されるが，ユーザーが任意に分割方法を指定することも可能である．

また，欠損値に関しては「欠損」カテゴリとして扱うことができるため，ある変数の欠損と他の変数の間に特別な関係がある場合，これを検出することが可能である．

カテゴリカル変数に対しても値の再カテゴリ化を適用することができる．たとえば，地域を表す変数が都道府県を単位とされて観測されているときに，これを「関東地方」「東海地方」といった形にまとめ直すことを指す．これは応用上の分類に合わせて行われることもあるが，カテゴリごとのサイズが小さくなりすぎることを防ぐ目的でも行われる．このため，Natto には観測数が一定の値を下回る値をまとめて `[Others]` として再カテゴリ化する機能が実装されている．

3·3 インターフェース

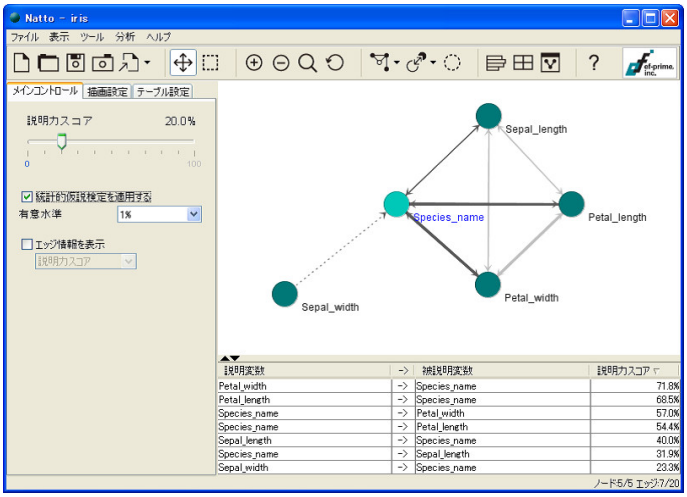

図 2 Natto メイン画面（グローバルモード；変数間の関係を探索するモード．ほかに相関ルール探索を行うローカルモードがある）

Natto のインターフェースは 図 2 のようになっている．画面右がメインウィンドウで，上部に変数間の関係を示したグラフが表示され，下部に同じ情報が表形式で表示される．画面上部のツールバーはグラフの拡大・縮小や平行移動，分析ツールの起動といった機能を提供する．画面左にはスライドバーなどのツールが配置され，これによってグラフの表示を調節することができる． メイン

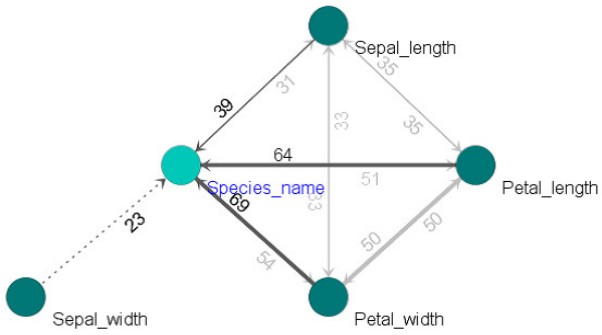

図 3 説明力スコアが 20% 以上の矢印を表示

ウィンドウには図 3 のように変数間の関係を表すグラフが表示されている．ここで各ノードは変数を表し，矢印は変数間の相関関係を表している．ここではアヤメの品種を表す変数 `Species_name` が選択されており，これに対して向かっている矢印が強調表示されている．矢印に付与された数値は説明力スコア（uncertainty coefficient）をパーセント表示したもので，これによって矢印の太さが変化する（ここでは 50% 以上のとき太線，30% 以上 50% 未満のとき細線，30% 未満のとき点線に設定されている）．これらはアヤメの種類に関する不確実性をどれだけ説明できるかを示し，言い換えればアヤメの種類

を予測するのに役立つ変数ほど高い値をとる．

図 3 より，Petal_width や Petal_length は Species_name に対して強い説明力を持っていることがわかる．また，Petal_width と Petal_length の間にもそれぞれ 50% の説明力があることから，この二変数の持つ情報はかなり重複していると考えられる．ここで，Sepal_width からは Species_name に対する矢印だけが表示されている．これは矢印の表示に関して説明力スコアの閾値を設定しているためで，画面左のスライダーによって調節することができる．また，変数間の独立性に関する仮説検定を行い，検定をパスした矢印のみを表示することもできる．ここでは説明力スコアの閾値を 20%，検定の有意水準を 1% としており，Sepal_width は Species_name 以外の変数と 20% 以上の説明力を持つほどの強い関係は持たないことを意味し，Species_name に関して他の 3 変数とは異なる情報を持っている可能性が示唆される．

実際，元の変数に対して主成分分析（相関行列による）を行った結果（図 4）を見ると，Sepal_width は他の変数との相関が低く，第二主成分に対する主成分負荷量が非常に高い（0.92）．他の 3 変数はそれぞれ相関係数が 0.82 から 0.96 と非常に高く，第一主成分に対する主成分負荷量が高い（0.52〜0.58）．つまり，これらの変数はひとつのグループを成していると考えることができ，これは図 3 において変数間の説明力スコアが互いに高かったことと一致する．

また，Petal_width と Petal_length の値によって品種の判別がうまくできそうであるのに対し，Sepal_length ではやや判別力が落ち，Sepal_width では図 4 中 2 と 3 で示された品種の判別は困難であることが伺われる．この傾向は図 3 における説明力スコアの傾向と一致している． このように，図 3 のグラフからデータセット全体

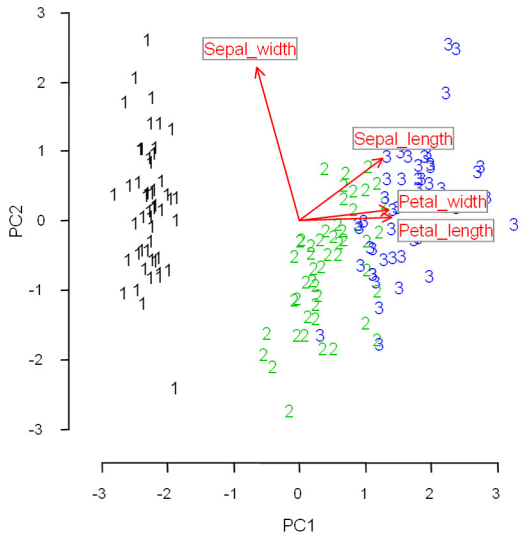

図 4 主成分バイプロット．数字はアヤメの品種を表す．

の性質を大まかに把握することができる．この方法の利点として，

（1） 変数を適切な大きさに分割することで，非線形の関係も捉えることができる

（2） 連続変数だけでなく，カテゴリカル変数や両者が混在した状況においても分析が可能である

ことがあげられる．同様の性質は CART [Breiman, et. al. 1984] をはじめとする決定木モデルにも当てはまるが，データ全体を一括して俯瞰することを目的とした場合には我々の方法がより適しているといえる．

3·4 分 析 ツ ー ル

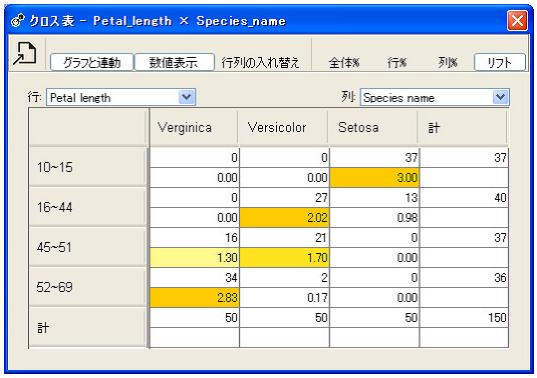

図 5 クロス集計表．相関ルールのリフトが高いものを色つきで表示している．

Natto には，変数間の関係をより詳細に分析するためのツールが用意されている．ひとつはクロス集計表であり，図 5 のようなインターフェースを持つ．変数のとる値同士の組み合わせについて，観測数 n_{ij} に加えて全体% (π_{ij})，行% $(\pi_{ij}/\pi_{i.})$，列% $(\pi_{ij}/\pi_{.j})$，または相関ルール $(X = x_i \rightarrow Y = y_j)$ のリフト $[\pi_{ij}/(\pi_{i.}\pi_{.j})]$ のいずれかが表示できる．これらの値によって表示色が自動的に変更され，確率の高い組み合わせや相関の高い組み合わせの有無を一目で確認することができる． また，相関

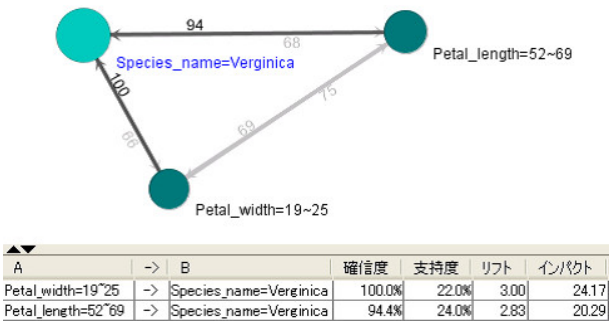

図 6 ローカルモードによる相関ルール探索．矢印はルールを示し，数値は確信度を表示している．下部はルールの表出力で，インパクトとは符号付 J-measure のことを指す．

ルール探索を行うローカルモード（図 6）も利用するこ

とができる．変数間の相関を探索するグローバルモードと同様のインターフェースを持ち，ノードは変数のとる値を，矢印はルールを示し，スライドバーによって符号付 J-measure，確信度，支持度，リフトに閾値を設定することができる．結果は表形式でも出力され，各指標によるソートや CSV 形式でのエクスポートが可能である．

3·5 規 模 耐 性

Natto では読み込んだデータをカテゴリ化してメモリに保管する．したがって変数の数やカテゴリ数が多いとメモリを大量に消費するが，サンプルサイズの増加に対しては比較的高いスケーラビリティを示す．また，あらかじめカテゴリの分割を決定してデータを読み込んでしまえば再度データソースにアクセスする必要がないため，データの探索を軽快に行うことが可能である．

ソフトウェアの規模耐性を検証するため，計算実験を行った．実験に用いたマシンの CPU は Intel Core Duo 1.66GHz（デュアルプロセッサ），メモリは 1GB で，Java VM に 512MB のメモリを割り当てた．実験用データとしては一様乱数による擬似データを作成し，グローバルモードの起動にかかる計算時間とメモリ使用量を測定した．実験データに対して計算時間とメモリ使用量を予測する回帰モデルを作成したところ，計算時間に関してはデータサイズ（サンプルサイズ × 変数の数）に対して 2 次の多項式が，メモリに関しては変数の数と分割数，およびその積に対して 2 次の多項式がよく当てはまった（決定係数はそれぞれ約 0.99 で，回帰係数はすべて有意水準 1%で有意）．

上記の結果から，利用可能なメモリサイズに対して変数の数に対する上限を推定することができる．しかし，ローカルモードの利用には更にメモリを消費すること，また変数の増加に従ってグラフの描画速度が低下することから，現実的には 2〜300 変数程度のデータに対して適用することが望ましい．

300 変数で分割数が 4（標準）の場合，メモリ使用量は 140MB 前後で，計算時間はサンプルサイズが 1000 のとき 11 秒，1 万のとき 48 秒，10 万のとき 6.5 分，100 万のときは 68 分で計算処理を完了した．各種指標を精度よく推定するために必要なサンプルサイズはデータの分布に依存するが，分析の目的が「データセット全体を俯瞰すること」であることを考えれば，大規模データに対しては数万件のサンプルを抽出して分析すれば良いだろう．

4. ま と め

本稿では変数間の相関関係について大域的・局所的な視点の両面から議論し，我々の開発したソフトウェア Natto を用いて探索的データ解析に応用した．また，既存の分析手法との比較を行い，その有用性が示された．

今後の発展の可能性として，3 変数以上の相互関係の評価や RDBMS との連携などが挙げられる．また理論的には，連続変数の分割に基づくエントロピー計算がどの程度本来の値を近似できるか，といった考察が課題といえる．

謝 辞

開発段階で有益なコメントをいただいた一橋大学の大上慎吾助教授，東京工業大学の下平英寿助教授ほか，ご助言いただいた皆様に感謝いたします．

◇ 参 考 文 献 ◇

[Agrawal and Srikant 1994] Agrawal, R. and Srikant, R.: Fast Algorithm for Mining Association Rules, Proceedings of the 20th VLDB Conference (1994)

[Borgelt and Kruse 2002] Borgelt, C. and Kruse, R.: Induction of Association Rules: Apriori Implementation, Proceedings of the 15th Conference on Computational Statistics (2002)

[Breiman, et. al. 1984] Breiman, et. al.: Classification and Regression Trees, Wadsworth (1984)

[Fisher 1936] Fisher, R. A.: The use of multiple measurements in taxonomic problems, Annals of Eugenics (1936)

[Goodman and Kruskal 1979] Goodman, L. A. and Kruskal, W. H.: Measures of Association for Cross Classifications, New York: Springer-Verlag (1979)

[Kullback 1959] Kullback, S.: Information Theory and Statistics, New York: Wiley (1959)

[O'Madadhain, et. al. (preprint)] O'MadadHain, et. al.: Analysis and Visualization of Network Data using JUNG, `http://jung.sourceforge.net/doc/` (preprint)

[Shannon 1948] Shannon, C. E.: A Mathematical Theory of Communication, The Bell System Technical Journal (1948)

[Smyth and Goodman 1992] Smith, P. and Goodman, R. M.: An Information Theoretic Approach to Rule Induction from Databases, IEEE Transactions on Knowledge and Data Engineering (1992)

[Theil 1970] Theil, H.: On the Estimation of Relationships Involving Quantitative Variables, American Journal of Sociology (1970)