確率的コンプレキシティを用いたルール学習による自由記述アンケート分析

李 航　　山西 健司
NEC 情報通信メディア研究本部

1　自由記述アンケート分析

アンケート調査は市場分析や顧客管理等を行う上で極めて重要である。特に自然言語で書かれた**自由記述アンケート**は多くの情報を含み、経営判断に必要な材料を提供すると考えられる。近年、情報技術の発達により Web 等を通じてアンケート調査が容易に行えるようになり、アンケートデータが大量に蓄積され、人手によるアンケート分析が労力的にもほとんど不可能になってきている。そこでコンピュータによる自由記述アンケート分析に対する期待が高まっている。

表1　アンケートデータの例

車種	ブランドイメージ
トヨタマーク II	大衆的な車だ
トヨタマーク II	乗りやすいね
…	…
日産セフィロー	人にやさしい感じ
日産セフィロー	ファミリーカーのイメージ
…	…

例えば、幾つかの車のブランドイメージについてアンケート調査を実施し、それぞれの車種のイメージに関する自由記述の回答が得られたとする(表1を参照)。その場合、車種とブランドイメージの間の関係性をコンピュータを用いて自動的に抽出することが分析の課題となる。その際、分析は以下の要件を満たさなければならないと我々は考える。

1. **分析対象に固有な特徴を表す言葉を正確に抽出すること**　それぞれの車種のイメージに関する回答に現れている**固有の言葉**を抽出し、それぞれの言葉が全体の中でどのぐらいの比率で出現しているかといった情報を数値的に捉えて提示すること。
2. **複数の分析対象の関係を抽出すること**　それぞれの車種のイメージだけでなく、全車種のイメージの間の関係を数量化して提示すること。
3. **分析結果を分りやすく視覚化すること**　1)、2）の分析結果をビジュアルに表現し、ユーザの理解を助けること。
4. **分析結果と元データの関係を分りやすく提示すること**　アンケート全体を要約した分析結果と元データとの対応づけを与えること。

一般に自由記述アンケートは表現がバラエティに富んでいて、データとして構造化されていないため、コンピュータでそれを扱うのは決して容易なことではない。現状では、自由記述の回答文を単語に分割し、単語レベルの分析を行うのが得策である。(単語レベルでもアンケートの内容を大まかに把握できる場合が実際には多い)。後述するように、従来ではこのような考えに基づき幾つかのアンケート分析システムが開発されたが、先に述べた全ての要件を満足するとは必ずしも言えないのが現状である。

我々は前記要件を全て満たす自由記述アンケート分析ツール **SurveyAnalyzer** を開発した。**SurveyAnalyzer** もやはり単語レベルでアンケート分析を行うものであるが、確率的コンプレキシティとよばれる統計尺度にに基づいて、先の要件1)、2）に対応しているところに大きな特徴をもっている。本稿では **SurveyAnalyzer** の原理と基本機能について概説する。

SurveyAnalyzer はルール型テキスト分類技術とコレスポンデンス分析技術に基づくアンケート分析を基本としている。

上記の例に対し、**SurveyAnalyzer** は、それぞれの車種をカテゴリ、車種のイメージに関する回答

をテキスト(実際、テキストを単語に分割する)と見なし、テキストとカテゴリの間の対応関係を表現するルールをアンケートデータから学習して表示する。

学習の際に確率的コンプレキシティをルール選択の規準にすることにより、それぞれの車種に固有な言葉を明確に抽出することができる。(→要件1)

また、SurveyAnalyzerは、複数の車種と、イメージ回答から抽出された複数のキーワードを確率変数とみなし、確率変数間のコレスポンデンス分析を行う。これによって車種とキーワードの相関関係を的確に分析することができる。(→要件2)

SurveyAnalyzerでは、上記分析結果を棒グラフとポジショニングマップを用いてビジュアルに表現し(→要件3)、また、検索機能を通じて分析結果と元となるデータとを常に対応できるようにしている(→要件4)。

2 SurveyAnalyzerの機能

SurveyAnalyzerでは、CSV形式のファイルで与えられたアンケートデータがあれば、「分析対象」と「自由記述回答」を指定することにより、直ちに分析を行うことができる。上記の例では、各車種が分析対象であり、ブランドイメージ回答が自由記述回答である。SurveyAnalyzerの機能を以下に挙げる。

- 製品や企業等の分析対象をカテゴリ、自由記述回答をテキスト(実際、テキストを単語に分割する)と見なし、テキストとカテゴリの間の確率的ルールを事例データから学習して表示する。ルールの種類としては、(新しい)自由記述回答を分類する目的に使われる**分類ルール**、自由記述回答からカテゴリへの想起関係を表す**相関ルール**、単純に頻度分析を行う**頻度ルール**の3通りを用意している。
- 複数の分析対象と、自由記述回答から抽出された複数のキーワードをそれぞれ確率変数とみなし、確率変数間のコレスポンデンス分析を行い、ポジショニングマップを出力する。
- ルールの学習におけるルール選択規準として統計的尺度「確率的コンプレキシティ」または「拡張型確率的コンプレキシティ」を用いる。
- ルールを棒グラフで、コレスポンデンス分析結果をポジションマップによって表現する。
- 検索機能を備え、ルールに現れたキーワードと、それが実際に出現した元データが直ちに対応できるので、キーワードが現れた実際の文脈を知ることができる。
- 操作性の優れたインターフェースを備え、初心者でも容易に分析することが出来る。
- ユーザが専門用語、不要語(例:「する」「なる」)、同義語(例:「計算機」と「コンピュータ」)などを登録できる。また「センスがよい/わるい」、「品質がよい/わるい」のような語句を登録し、まとめて扱うことができる。さらに、肯定表現と否定表現も自動的に区別して扱う。
- 例えば1000件程度のアンケートデータを10秒以内で高速処理することができる、

3 ルールによる分析

3.1 分類ルールと相関ルール

分類ルールとは、一連のIF-Then-Elseタイプのルールのことである。例えば、SurveyAnalyzerは表1のアンケートデータから図1に示す分類ルールを学習して出力する。

図1 分類ルールの例

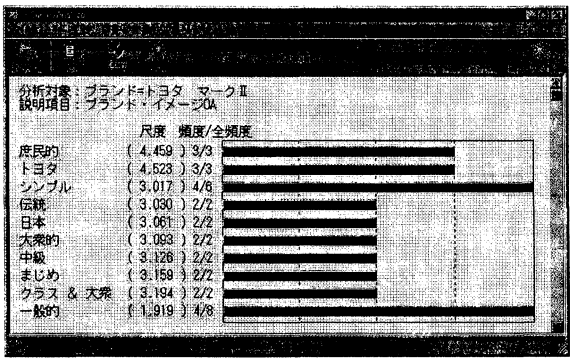
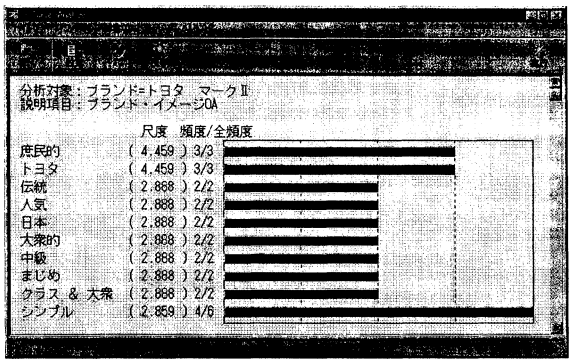


図1では、第一行のルールは「庶民的」という単語が全車種のイメージの回答に3回現われ、そ

の内３回とも「トヨタマーク II」のイメージの回答に現われたことを表す。第二行のルールは、**上記該当回答(「庶民的」が現われた回答)を除いたものの中で**、「トヨタ」という単語が全車種のイメージの回答に３回現われ、その内３回とも「トヨタマーク II」のイメージの回答に現われたことを表す。以下続く。ルールは単語の出現の連言を条件にするもの(例えば、第九行のルール「クラス＆大衆」)も含む。このような分類ルールは**確率的決定リスト**とよばれる[2,4]。

図の中では「全頻度」は単語の全車種のイメージ回答に現われる度数を表し、「頻度」は単語の分析対象の車種のイメージの回答に現われる度数を表す。また、「尺度」はルールを学習する際に用いた統計的尺度による「情報利得」の計算値で([2]を参照)、棒グラフの長さは頻度の大きさを表す。

図２　相関ルールの例



一方、相関ルールとは、IF-Else-Or タイプのルールである。例えば、**SurveyAnalyzer** は表１のアンケートデータから図２に示す相関ルールを出力する。

図２では、第一行のルールは「庶民的」という単語が全車種のイメージの回答に３回現われ、その内３回とも「トヨタマーク II」のイメージの回答に現われたことを表す。第二行のルールは、**引続き全回答中で**、「トヨタ」という単語が全車種のイメージの回答に３回現われ、その内３回とも「トヨタマーク II」のイメージの回答に現われたことを表す。以下続く。

分類ルールは、カテゴリ（車種）が未知である自由回答データにカテゴリを振分ける（＝分類する）ためのルールであるのに対して、相関ルールは、カテゴリを想起するキーワードを強い順に列挙するルールである。

3.2 ルールによる特徴抽出

分類ルールあるいは相関ルールは分析対象の固有の特徴を抽出したものである。これは単純に頻度の高い単語を列挙して得られる頻度ルールとの比較で明らかになる。

図３　頻度ルールの例

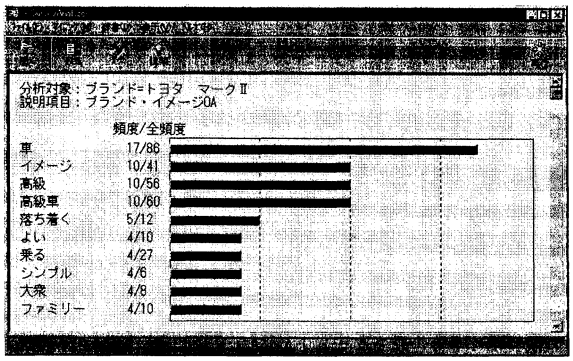


図３はトヨタマーク II のイメージの回答に現われる頻度の高い上位１０単語である。そこには「車」「イメージ」等どの車種にも現われる単語が出ている。従って、それらの単語がマーク II の特徴を表すとはいえない。つまり、一般には出現頻度の高い単語からなるルールで分析対象の固有の特徴を表すことができないのである。

図４　分類ルールの例

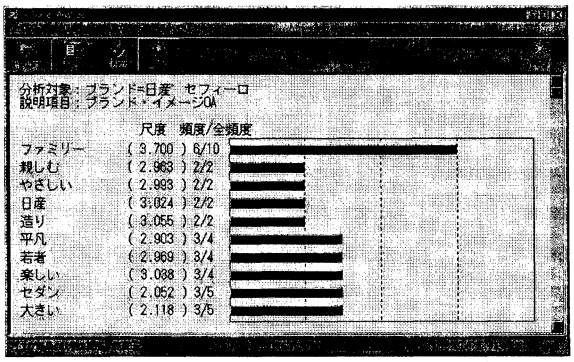


図２によれば、トヨタマーク II のブランドイメ

ージに関する分類ルールの第1行に「庶民的」という単語が3/3の割合で現われている。一方、図4は日産セフィーロのブランドイメージに関する分類ルールであるが、第1行に「ファミリー」という単語が6/10の割合で現われている。従って、「庶民的」がマークIIの特徴を表す言葉で、「ファミリー」がセフィローの特徴を表す言葉であると言える。ある車種に偏って出現している単語がその車種の特徴を表していると考えることができる。

分類ルールと相関ルールのいずれも分析対象に偏って出現する特徴的な単語を抽出することができる。特に、分類ルールは分析対象の特徴の要約に適しており、相関ルールは分析対象の特徴の発見に適しているといえる。

3.3 統計的尺度

SurveyAnalyzerは統計的尺度**確率的コンプレキシティ**と**拡張型確率的コンプレキシティ**をルール選択の基準としてルールを生成することを特徴としている。以下、これらの統計的尺度について解説する。

確率的コンプレキシティ(Stochastic Complexity、SCと略記)[3]は与えられたデータに含まれる情報の量を表す尺度である。これは、与えられた確率モデルを用いてデータを符号化するための最短符号長(或は、記述長)として、モデルに相対的に定義される。与えられたデータに対して最も小さいSCを実現出来る確率モデルを最良なモデルとみなす原理を**MDL原理**(Minimum Description Length Principle、記述長最小原理)と呼ぶ。直観的には、データに確率モデルをあてはめる際に、自らの長さも含めてデータを最も短く記述できるような確率モデルをあてはめようとする原理である。

確率的コンプレキシティは、モデルの形式を確率モデルに、情報の測る尺度を符号長に限定したときに定義されたものである。一方、**拡張型確率的コンプレキシティ**(Extended Stochastic Complexity、ESC)[5]とは、同様に与えられたデータに含まれる情報の量を測定する尺度であるが、モデルを確率モデルに限らず一般の関数に置き換え、情報の測り方も一般の損失関数(例えば、分類誤り数)を許すことで、確率的コンプレキシティを一般化したものである。

MDL原理は、真の確率分布を高い精度で推定できるといったよい性質をもつことが理論的に証明されている[1、3、4]。また、未知のデータに対して、予測(分類)誤りを最少にすることを目的とした場合には、ESC最小原理はMDL原理より分類誤り率を下げることができる[5]。

3.4 ルールの学習アルゴリズム

上記統計的尺度を用いた分類ルールの学習アルゴリズムの詳細は[2]を参考にされたい。

参考文献

[1] Andrew R. Barron and Tomas M. Cover, Minimum complexity density estimation, IEEE Transactions on Information Theory, 37(4):1034-1054, 1991.

[2] Hang Li and Kenji Yamanishi, Text classification using ESC-based stochastic decision lists, Proceedings of the 8th International Conference on Information and Knowledge Management (ACM-CIKM'99), 122-130, 1999.

[3] Jorma Rissanen, Fisher information and stochastic complexity, IEEE Transaction on Information Theory, 42(1):40-47, 1996.

[4] Kenji Yamanishi, A learning criterion for stochastic rules, Machine Learning, 9:165-203, 1992.

[5] Kenji Yamanishi, A decision-theoretic extension of stochastic complexity and its applications to learning, IEEE Transactions on Information Theory, 44(4):1424-1439, 1998.