知能と複雑系 116－1
（1999．5．24）

Webテキストから獲得した制約型確率知識を扱う超空間推論法

藤本 和則　　松澤 和光

NTT コミュニケーション科学基礎研究所

〒619-0237 京都府相楽郡精華町光台2 - 4

tel: 0774-93-5359

E-mail: {fujimoto, matuzawa}@cslab.kecl.ntt.co.jp

概要： 我々は、インターネットユーザに、統計的決定の見地からアドバイスを与えるDSIUシステムの研究を進めている(DSIU：*Decision Support for Internet Users*)。本稿では、DSIUシステムのもつ機能のうち、ネット上の記述文から獲得された知識を使って推論を行う**超空間推論法**について述べる。この超空間推論法は、(1) ネット上の記述文を主観確率についての制約式に変換する**主観確率解釈**と、(2) 得られた制約式の集合から主観確率を一意に定める**主観確率確定**とからなる。本稿では、このうち、特に後者について、その定式化を行う。超空間推論法の考え方を述べるにあたっては、ネット上の電子製品に関する記事を例として取り上げる。

Hyper-Space Reasoning for handling probability constraints acquired from statements on the Internet

Kazunori FUJIMOTO and Kazumitsu MATSUZAWA

NTT Communication Science Laboratories

2-4 Hikaridai Seika-cho Soraku-gun Kyoto 6190237 Japan.

tel: +81-774-93-5359

E-mail: {fujimoto, matuzawa}@cslab.kecl.ntt.co.jp

Abstract： This paper presents a principle that elicits subjective probabilities from statements on the Internet. This principle consists of (1) *subjective probability interpretation* that transforms statements on the Internet into constraints on the subjective probabilities of humans who describe the statements, and (2) *subjective probability determination* that elicits subjective probabilities from the set of the constraints. This principle is introduced by referring to an example of articles for electric products on the Internet.

1 はじめに

近年、インターネットの普及により、人間は世界各国に散在する最新の情報を瞬時に取得できるようになった。こうした状況にあわせて、ネットから有効な情報を集めて、人間の活動に役立てる研究が始められている [2, 5]。こうした研究に多くの努力が払われているものの、ネット情報に基づく判断は依然難しいといえる。この原因としては、ネット情報が膨大であったり (大量の情報から該当するものを集めるため)、あまり知られていなかったり (次々と新しい情報へ更新されるため) することがあげられる。こうしたネット情報の性質が、それらから適切な判断を下す作業を困難にしている。そこで、我々は、ユーザの判断に対して統計的決定の見地からアドバイスを与えるシステムを DSIU システム[1]と呼び、その実現を目指して研究を進めている [6, 13]。

統計的決定を行うためには、基本的に「決定の結果生じる状況を確率的に予想する確率知識」と「生じた状況の良し悪しを定量的に評価する効用データ」の二つの獲得が必要となる。このうち後者は、ユーザの関心ある状況のみに限定して獲得すれば良いので、数量化にあたっての困難は多少あるかもしれないが、従来から提案されている様々な方法 [11] に基づいて獲得できるであろう。これに対して、前者は、ネット情報、すなわち、幅広い内容を含み、刻一刻と新しくなるような情報を扱えるものであることが要求される。したがって、人間が直接に確率知識を用意するような方法では、膨大な量の収集、頻繁な更新に莫大なコストが見込まれ、現実的とはいえない。我々の目指すシステムの実現には、こうした確率知識をいかにして獲得するかが重要となる。

こうした問題に対して、我々は「ネット情報を扱って判断するための知識は、それ自体ネットに存在するのではないか」と考えた。例えば、ネットで提供される製品カタログや、各機種についてのディスカッションなどには「ある仕様が性能を良くする」といった記述が見られる。こうした文から、仕様と性能に関しての依存関係が獲得できるかもしれない。ネット上に存在するこうした文は、それ自体豊富な内容を含んでおり、また、次々に最新のものに更新されていく。こうした文から自動的に確率知識を獲得できれば、様々なネット情報を扱う確率知識ベースを構築することができるだろう。以上のような考えに基づき、我々は、ネット上の文から確率知識を自動獲得する方法の研究を進めている。

我々は、これまでに、ネット上の記述文から確率知識を獲得するための基本原理を提案した [13]。そして、こうした獲得を自動で行う方法を提案し、実際のネット上の記事を用いて、その有効性を確認した [12]。こうした自動獲得に基づく方法では、主にテキスト解析の誤りから、誤った知識の獲得を避けられない。そこで、本稿では、先に提案した知識獲得のための基本原理を誤りの含まれる知識を扱えるように拡張する。本稿では、この拡張された基本原理を**超空間推論法**と呼ぶ。この超空間推論法は、(1) ネット上の記述文を主観確率についての制約式に変換する**主観確率解釈**と、(2) 得られた制約式の集合から主観確率を一意に定める**主観確率確定**とからなる。本稿では、まず、2 章で主観確率解釈の考え方を述べた後、3 章で主観確率確定の定式化を行う。

2 主観確率解釈

本章では、ネット上の記述文を主観確率についての制約式に変換する主観確率解釈について述べる。より具体的な論述のため、ネット上の記述文として、デジタルカメラの製品記事を取り上げる。そして、この記事から、あるデジタルカメラの機種について、その仕様 (例えば、焦点方式、レンズ) から特性 (画質) を予測する確率知識を獲得する問題を考える。

主観確率解釈では、まず、ネット上の記述文を「それを記述した人間のもつ主観確率についての制約式」として解釈する。例えば、「仕様 s が画質を良くする」という記述は、それを記述した人間の主観確率について「仕様 s を採用すると、機種の画質が良い確率は大きくなる」という情報を与えていると解釈するのは妥当であろう。こうした解釈は、次のように定式化することができる。

$$\Pr(\text{画質} = \text{良い} | \text{仕様} = s) > \Pr(\text{画質} = \text{良い})$$

以上のように解釈することにより、ネット上の記述文は「確率パラメータのとりうる数値 (確率値) の範囲を小さくする情報」として利用することが可能と

[1]ここに、DSIU は“デシュウ”と読み、*Decision Support for Internet Users* の略である。

なる。

主観確率の獲得については、これまで、適切な表現 [8, 9]、言葉との関係 [1, 3] の面から様々な研究が行われてきた。しかしながら、これらの研究では、ネット上に展開された記述文と主観確率との関係については調べられていない。我々の研究では、ネット上の記述文から主観確率を獲得することを目指すものである。したがって、(1) ネット上のどのような情報を利用できるか、(2) それをどのような形式の情報として解釈するのが妥当か、についての研究が重要となる。こうした視点からは、従来の主観確率の研究は、その研究対象をネット上の記述文を含めたものに拡張される必要がある。

Pearl は、人間の記述した知識として論理式に焦点をしぼり、その確率的解釈を研究した [10]。また、Goldszmidt は、believable, unlikely などの言葉 (linguistic quantifiers) も扱えるように拡張した [7]。こうした考えをネット上の記述文に拡張するためには、言葉のみでなく、その他の多くの情報を考慮する必要があるだろう。例えば、文章中の記述の位置や、記述に用いられた文字のサイズあるいは、記述の時期などの情報も、主観確率を獲得するにあたって重要となるかもしれない。より具体的な例として、製品カタログ中に二つの文「仕様 s_1 は画質を良くする」、「仕様 s_2 は画質を良くする」が書かれていた場合を考える。これらの文が独立に書かれている場合、仕様 s_1 と s_2 の間には、何の違いもみいだせない。しかしながら、前者の方が後者に比べより大きな文字サイズが用いられて書かれていた場合は状況が異なる。この場合、それを記述した人間は、前者の方を強調したかったと解釈できるので、仕様 s_1 の方が s_2 に比べ画質により大きな効果を与えると解釈するのは妥当であろう。こうした解釈は、次のように定式化できる。

$$\Pr(\text{画質} = \text{良い}|\text{仕様} = s_1) > \Pr(\text{画質} = \text{良い}|\text{仕様} = s_2)$$

このように、主観確率解釈は、ネット上の様々な情報から主観確率についての制約式を引き出して、それらを確率知識の獲得に役立てようとするものである。

この主観確率解釈での情報抽出は、テキスト解析に基づいて行われる。こうした解析には、比較的難しいものと容易なものとがある。例えば、自然言語文の係受け解析は、HTML のタグを利用した解析に比べて、表現の多様さから比較的困難であるといえる。解析の困難な抽出に基づいて得られた制約式は、“得られた制約式が正しい確率が低い”という意味で、信頼性の低い制約式であるといえる。主観確率の導出にあたって、こうした信頼性を知ることができれば、より精度の高い導出ができるかもしれない。そこで、主観確率解釈では、各制約式に「その制約式が正しい確率」をその信頼性の程度として与える。この信頼性の程度を**解釈信頼度**と呼ぶ。この解釈信頼度の値を決めるにあたっては、実際の解析誤りの統計情報から定めることもできるし、“記述文解析の難しさの程度”という人間の主観に基づいて決めることもできる。以上のように、主観確率解釈は、テキスト解析に基づいて、ネット上の記述文を“解釈信頼度を伴った確率制約式の集合”に変換するものである。

3 主観確率確定

本章では、主観確率解釈により得られた制約式の集合から主観確率を導く主観確率確定の定式化を行う。我々は、こうした確定法に要求される性質として、次の二つを抽出した。

- 様々な形式の制約式を扱える：ネット上の多様な情報を表した制約式は、様々な形式をとることが予想される。こうした様々な形式の制約式のもとで、主観確率を導出する必要があるため。
- 制約式の矛盾を扱える：ネットから集められた制約式は、誤った制約式の混入により矛盾をもつことがある。こうした矛盾のある制約式のもとで、主観確率を導出する必要があるため。

本章では、これら二つの扱いを実現する主観確率確定法について述べる。まず、3.1 節で、様々な形式をとる確率制約式の統一的な扱いを可能にする分布超空間 [4] の考えを説明する。主観確率確定は、この分布超空間のモデルを採用することにより、様々な形式の確率制約式の扱いが可能となる。そして、3.2 節で、主観確率を導くにあたって、分布超空間のモデル上で解釈信頼度を総合的に判断する主観確率確定の定式化を行う。これにより、制約式に誤りがあっても、各制約式の解釈信頼度に基づいて、適

切な主観確率を導くことが可能となる。3.3節では、製品カタログの記事から仕様と特性についての主観確率を獲得する問題を取り上げ、主観確率確定の例示を行う。なお、本章で扱う確率変数は全て離散型とする(連続型のものは適当な量子化により離散型にできるとする)。

3.1 分布超空間

問題領域中の全ての確率変数についての同時確率に着目し、この同時確率を軸として張った実数値空間を分布超空間という。例として、二つの確率変数、画質 $Q \in \{q_1, q_2\}$、レンズ $L \in \{l_1, l_2, l_3, l_4\}$ を取り上げる(ここに q_1, q_2 は'良い'、'普通'などの画質の評価、$l_i,\ i = 1, \ldots 4$ は異なるレンズをそれぞれ表す)。こうした二つの確率変数からなる問題において、同時に起こり得る全ての場合

$$(q_1l_1, q_1l_2, q_1l_3, q_1l_4, q_2l_1, q_2l_2, q_2l_3, q_2l_4) \qquad (1)$$

のそれぞれを構成割当と呼ぶ(ここに、q_il_j は $Q = q_i,\ L = l_j$ が同時に起こる事象を指す)。そして、この構成割当の上への確率によって張られる空間を分布超空間と呼ぶ。この構成割当の上への確率を基底確率と呼ぶ。ここでは、リスト(1)に対応した八つの基底確率をそれぞれ $x_1, x_2, \ldots, x_8$ と書くことにする。こうした基底確率に基づいて、例えば、確率公理からくる制約は、それぞれが1以下の非負実数で、その総和が1(リスト(1)の成分は互いに排反で、全てを尽くしているため)という制約式として書くことができる。

$$0 \leq x_i \leq 1,\ i = 1, \ldots 8,$$
$$x_1 + x_2 + \cdots + x_8 = 1$$

また、例えば、

$$\Pr(Q = q_1 | L = l_1) > \Pr(Q = q_1)$$

という確率表現は、ベイズの規則と簡単な確率規則に従って、基底確率についての次のような制約式に変形して書くことができる。

$$\frac{x_1}{x_1 + x_5} > x_1 + x_2 + x_3 + x_4$$

同様に、確率区間や質的影響などの表現も、基底確率からなる式として書くことができる[4]。以上に述べたように、様々な形式をもつ制約式は、基底確率により張られる分布超空間上の制約式として統一的に扱うことができる。主観確率確定においては、様々な形式の制約式を扱えるようにするため、この分布超空間のモデルを採用する。

3.2 定式化

主観確率解釈により、主観確率分布 Pr に関して、いくつかの制約式が得られる。こうした制約式の解釈信頼度は、以下のように、分布超空間の点について確率分布 $\Pr^\star$ を与える。

1. 各制約は、分布超空間について、その制約の充足する領域に確率を与える。この確率は、解釈信頼度により定められる。
2. この分布超空間上の確率から、分布超空間の各点について、その点が真である確率が計算される。

一方、3.1節に述べたように、分布超空間の各点は、ただ一つの Pr を与える。結果として、解釈信頼度から、Pr の期待分布が導かれる。主観確率確定は、この Pr の期待分布をもって、Pr の推定値とするものである。

以下では、こうした主観確率確定の定式化を行う。

- 確率変数の集合を V、また V についてのすべての構成割当の集合を W とする。また、この W の要素数(すなわち構成割当の数)を k とする。
- 構成割当の集合 W から得られる分布超空間の量子化レベルを q とする(すなわち、分布超空間の各軸を $1/q$ きざみに量子化する)。結果として、分布超空間は、q^k 個の点を含むことになる。この点を量子点と呼ぶ(量子化レベル q は、分布超空間の状態を知るに十分な大きさであると仮定する)。
- 分布超空間を $\mathcal{S} = \{e_1, \ldots, e_{q^k}\}$ とする。ここに、$e_i,\ i = 1, \ldots, q^k$ は、分布超空間 $\mathcal{S}$ の量子点を表す。

ここで、主観確率分布 Pr についての制約式の集合 $C = \{C_1, \cdots, C_n\},\ n \geq 1$ を考える。ここに、各制約 C_i の解釈信頼度をそれぞれ α_i とする。制約式の矛盾を避けるため、各解釈信頼度は、0から1の間の値をとり、0, 1そのものをとらないと仮定する。

- 制約式 C_i を充足する量子点の集合 ($\in \mathcal{S}$) を M_i, $i = 1 \ldots n$ とする。また、M の補集合と要素数をそれぞれ $\overline{M}$ と $|M|$ とする。

制約式 ($\in C$) の解釈信頼度に基づいて、分布超空間 $\mathcal{S}$ 上の確率分布 $\Pr^\star$ は次のように書ける (但し、各制約式の解釈信頼度の独立性を仮定した)。

$$\Pr^\star(e) = \frac{\prod_{i=1}^{n} \delta_i(e)}{\left| \bigcap_{i=1}^{n} \eta_i(e) \right|} \qquad (2)$$

ここに、e は "量子点 e は真" という命題、関数 δ_i は、

$$\delta_i(e) = \begin{cases} \alpha_i & \text{if } e \in M_i \\ 1 - \alpha_i & \text{if } e \notin M_i, \end{cases}$$

関数 η_i は、

$$\eta_i(e) = \begin{cases} M_i & \text{if } e \in M_i \\ \overline{M}_i & \text{if } e \notin M_i \end{cases}$$

をそれぞれ表す。量子化レベル q が十分に大きいときは、式 (2) に分母は、空集合とならないことに注意されたい。以上から、確率分布 Pr の期待確率は次のように書ける。

$$\tilde{\Pr}(t) = \sum_{e \in \mathcal{S}} [\Pr(t)]_e \Pr^\star(e) \qquad (3)$$

ここに、t は、V 中の確率変数へのある割当を表す ("$Q = q_1 | L = l_1$" などの条件付表現も許される)。また、確率 $[\Pr(t)]_e$ は、量子点 e から導かれた確率 $\Pr(t)$ を表す。式 (3) を用いることによって、主観確率確定は、主観確率分布 Pr の推定値 $\tilde{\Pr}$ をその期待分布として導くことができる[2]。

3.3 例題

本節では、3.2 節に述べた主観確率確定について、具体例をもとに実際に主観確率の計算を行う。例題としては、デジタルカメラについて、採用するレンズの種類から、その機種の画質の善し悪しを推測する問題を取り上げる。確率変数としては、画質 $Q \in \{q_1, q_2\}$、レンズ $L \in \{l_1, l_2, l_3, l_4\}$ の 2 つを用いる。(q_1, q_2 はそれぞれ '良い', '悪い' という評価、$l_i, i = 1, \ldots, 4$ は異なる 4 つのレンズをそれぞれ表す)。制約式としては、次の二つを用いた。

$$\Pr(Q = q_1 | L = l_1) > \Pr(Q = q_1) : [\alpha_1] \qquad (4)$$

$$\Pr(Q = q_1 | L = l_1) > Pr(Q = q_1 | L = l_2) : [\alpha_2] \qquad (5)$$

式 (4), (5) に見られるように、各式の解釈信頼度は、それぞれ α_1, α_2 で表した。式 (4), (5) は、それぞれ $\alpha_1, \alpha_2 \geq 0.5$ のときには、確率 $\Pr(Q = q_1 | L = l_1)$ が大きい値を取ることを主張する。逆に、$\alpha_1, \alpha_2 < 0.5$ のときには、確率 $\Pr(Q = q_1 | L = l_1)$ が小さい値を取ることを主張する。ここに、確率 $\Pr(Q = q_1 | L = l_1)$ は、"l_1 を採用した機種は画質が良い確率" を表す。分布超空間を構成するにあたっては、計算時間の削減のため、10^5 個の点をランダムサンプリングして分布超空間を構成した。以上の条件のもとに、確率 $\Pr(Q = q_1 | L = l_1)$ について主観確率確定を行った。この結果を図 1に示す。

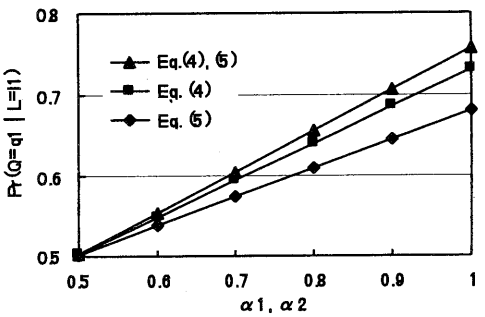

図 1: 確定結果

図 1において、横軸は解釈信頼度の値、縦軸は確率 $\Pr(Q = q_1 | L = l_1)$ の確定値をそれぞれ表す。解釈信頼度としては、α_1, α_2 を 0.5 から 0.1 きざみで 1 まで動かして確定を行った。四角と菱形の点は、それぞれ、式 (4) と (5) から確定された値を示す。これに対して、三角の点は、これら二つの式を合わせて確定された値を示す。なお、解釈信頼度が 0.5 以下の場合の結果は、解釈信頼度が 0.5 の点についての対称な直線として得られる。

図 1から、制約式の解釈信頼度が大きくなるほど、確定された確率値が大きくなっていることがわかる。

[2] この確定は、人間の頭の中に確定的な主観確率値が存在することを主張するものではなく、値の確定により、以降の統計的決定計算を容易にすることをねらったものである。

これは、確率 $\Pr(Q = q_1|L = l_1)$ が大きいことを支持する制約の信頼性が増すと、その確定値が大きくなることを意味する。このように、主観確率確定は、解釈信頼度の大きな制約式ほど、制約式としての効果を大きくする計算法となっていることがわかる。また、図1から、解釈信頼度が一定の条件のもとでは、二つの式を合わせた場合の確定値は、それぞれの式単独の場合の確定値より、大きくなっていることがわかる。これは、確率 $\Pr(Q = q_1|L = l_1)$ が大きいことを支持する制約式が増えるほど、その確定値が大きくなっていることを意味する。このように、主観確率確定は、確率値の大きさについて同様に支持する制約式が増えるほど、その支持が強く反映される計算法となっていることがわかる。以上のように、主観確率確定では、解釈信頼度を総合的に判断して、適切な主観確率を導くことができる。

4 おわりに

本稿では、ネット上の記述文から確率知識を構成する超空間推論法について述べた。我々は、デジタルカメラの製品記事を例に、実際に、信頼度付の制約式を獲得する研究を進めている。今後は、こうして獲得された制約式を用いて、実際に、どれくらい正確な主観確率を導出できるかを調べていく予定である。

謝辞

本研究を進めるにあたって、熱心に討論頂きました、村松 純 氏、向内 隆文氏 (NTT コミュニケーション科学基礎研究所) に深く感謝します。

参考文献

[1] Ruth Beyth-Marom. How probable is probable? a numerical translation of verbal probability expressions. *Journal of Forecasting*, Vol. 1, pp. 257–269, 1982.

[2] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. Learning to extract symbolic knowledge from world wide web. In *AAAI–98*. The AAAI Press, 1998.

[3] Marek Druzdzel. Verbal uncertainty expressions: Literature review. CMU-EPP-1990-03-02, Carnegie Mellon University, 1989.

[4] Marek J. Druzdzel and Linda C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pp. 141–148, Montreal, Quebec, Canada, 1995.

[5] Dieter Fensel, Stefan Decker, Michael Erdmann, and Rudi Studer. Ontobroker: The very high idea. In *Proceedings of the Eleventh International Flairs Conference (FLAIRS–98)*, 1998.

[6] Kazunori Fujimoto and Kazumitsu Matsuzawa. Intelligent systems using web-pages as knowledge base for statistical decision making. to appear in *New Generation Computing*, Vol. 17, No. 4, 1999.

[7] Moisés Goldszmidt and Judea Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, Vol. 84, No. 1, pp. 57–112, 1996.

[8] David Heckerman and Holly Jimison. A bayesian perspective on confidence. In L. N. Kanal, T. S. Levitt, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pp. 149–160. Elsevier Science Publisher, 1989.

[9] Gerhard Paaß. Second order probabilities for uncertain and conflicting evidence. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pp. 447–456. Elsevier Science Publisher, 1991.

[10] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[11] Detlof Von Winterfeldt and Ward Edwards. *Decision analysis and behavioral research*. Cambridge University Press, 1986.

[12] 賀沢秀人, 藤本和則, 松澤和光. Web テキストを知識ベースとして用いる推論システムの提案：テキストからの知識獲得方式を中心に. 人工知能学会研究会資料 SIG-KBS-9803-9, pp. 49–54, 1999.

[13] 藤本和則, 松澤和光. インターネット上の記述文から確率知識を構成する一手法 : 構成の基本原理を中心に. 情報学シンポジウム, pp. 129–136, 1998.