Web文書を使った主観確率の獲得: 統計的意思決定の視点から

Use of Web-Pages to Acquire Subjective Probabilities: From A Viewpoint of Statisitical Decision Making

藤本 和則
Kazunori FUJIMOTO

有限会社フジモト・リサーチパーク 人工知能研究所
Artificial Intelligence Laboratory, Fujimoto Research Park Co., LTD.

We research a mathematical model for decision making processes where the knowledge for the decision is constructed automatically from information on the Internet. In this model, rational decisions are made based on statistical decision making theory and five elements for the theory are acquired from the decision maker and information on the Internet. This paper discusses one of the knowledge acquisition processes where probability distributions, which are used to calculate expected utilities, are acquired from Web-pages on the Internet.

1. はじめに

近年、自然言語処理や知識獲得技術 [15, 18] の進歩により、インターネット上の情報から様々な知識をある程度の精度で自動獲得できるようになってきた。こうした状況に合わせて、ユーザの意思決定に必要な知識をネット上の情報から自動構成する研究が始められている [4, 12]。本研究では、特に、統計的意思決定 (例えば [14]) に必要な「代替案の集合、データの集合、確率分布」について、これらをネット情報から自動獲得する意思決定支援の研究を進めている [16, 17]。

本研究では、これまでに、「ネット情報から自動獲得した知識で意思決定できるか」という基本問題について数理的な知見を得るため、意思決定系の基本モデルを示した [17]。この系を DSIU と呼ぶ。ここに、DSIU は *Decision Support for Internet Users* の略である。図 1 に DSIU の基本モデルを示す。DSIU の基本モデルは、意思決定の過程だけでなく、意思決定に必要な知識を獲得する過程 (図中の知識獲得システム) を含むところに特徴がある。こうした知識獲得の過程をモデル化することにより、知識源の充実さの程度、知識獲得の精度といった要因を取り込んで、意思決定を議論することが可能となる。

本稿では、DSIU の知識獲得過程のうち、そのモデル化が最も困難となる確率分布の獲得について述べる。まず、確率分布の獲得に役立つ知識として、Web 文書のもつ主観的な知識 (主観評価知識と呼ぶ) について説明し、この表現法を具体的に示す。そして、主観評価知識を使った意思決定を、不正確性、断片性、矛盾性の 3 つから特徴づけ、これらの不完全性のもとに意思決定を行うための技術について論じる。

連絡先: 〒 532-0004 大阪市淀川区西宮原 2-7-59-1204,
有限会社フジモト・リサーチパーク,
TEL & FAX: 06-6397-3057,
E-mail: frp@palette.plala.or.jp.

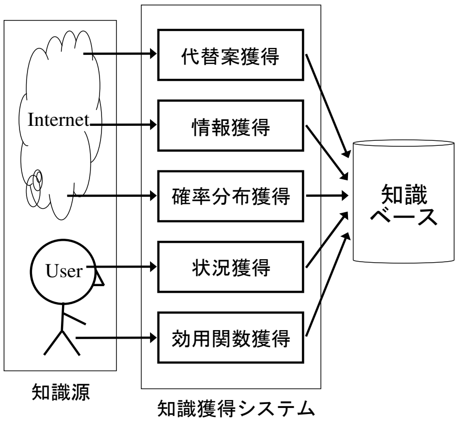

図 1: 意思決定系の基本モデル

2. 確率分布獲得のモデル

計算機上で合理的な意思決定を行うため、本研究では、Web 文書のもつ主観的な知識を使って確率分布を特徴付ける。この確率分布は、決定によりもたらされる結果を予測するためのものである。例えば、デジタルカメラに関する

「211 万画素により超高画質。」 (1)

という Web 文書中の記述は、あらっぽく言えば、画質 Q と撮像素子 S について、

「$\Pr(Q = \mathrm{good}|S = 211)$ が大きい値をとる」

というように確率分布を特徴付ける。こうした知識は、仕様から画質を予測するものであり、購入する機種を決めるという意思決定に利用できる。このように、人間の主観から確率分布を特徴付ける知識を主観評価知識と呼ぶ。本章では、まず、2.1 節で、主観評価知識の知識表現について述べる。そして、2.2 節で、主観評価知識を使った意思決定について、その不完全性を中心に述べる。

2.1 知識表現

主観評価知識は、Web 文書中の質的情報、程度情報、そして比較情報の 3 種類の情報を使って "確率分布についての制約式" として獲得される。質的情報は、どの確率値についての知識か指定する情報であり、例えば、知識 (1) については、$\Pr(Q = \text{good}|S = 2.11\text{M})$ を指定する情報である。程度情報は、確率値の大きさの程度を指定する情報であり、例えば、

$$0.7 \le \Pr(Q = \text{good}|S = 2.11\text{M}) < 0.8 \qquad (2)$$

といった区間を表す情報である。比較情報は、複数の確率値の大小関係を指定する情報であり、例えば、

$$\Pr(Q = \text{good}|S = 2.11\text{M}) \\ \ge \Pr(Q = \text{good}|S = 1.44\text{M}) \qquad (3)$$

といった比較を表す情報である。

以上のように、本研究では、確率分布についての制約表現からなる表現空間を考え、Web 文書中の質的、程度、比較情報から表現空間へのマッピングとして主観評価知識を獲得する。

2.2 知識ベースの不完全性

意思決定にあたっては、まず、Web 文書から主観評価知識を獲得して知識ベースを構成し、こうした知識ベースを使って意思決定が行われる。この過程を図 2 に示す。

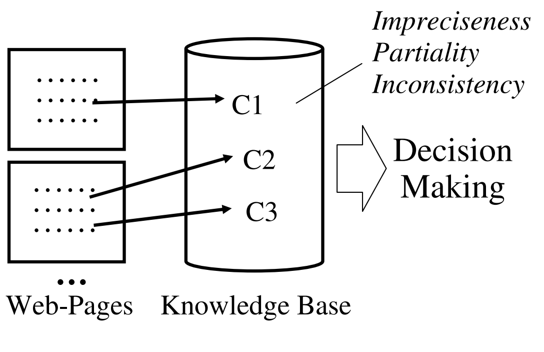

図 2: 知識獲得と意思決定の過程

こうした意思決定の過程において、知識ベースは、多くの場合、不正確で、断片的で、矛盾がありえるという特徴をもつ。

- 不正確性 ($\boldsymbol{Impreciseness}$) 各主観評価知識は、確率値を一意に決める情報をもつとは限らない。例えば、知識 (2)、(3) は、$\Pr(Q = \text{good}|S = 2.11\text{M})$ の値を正確に (例えば 0.8 のように) 定めない。
- 断片性 ($\boldsymbol{Partiality}$) 主観評価知識は、全ての事象について獲得されるとは限らない。例えば、知識 (2)、(3) は、画素数が 81 万画素である機種に関しては言及しない。
- 矛盾性 ($\boldsymbol{Inconsistency}$) 主観評価知識は、お互いに矛盾することがある。例えば、評価の辛い知識提供者が、211 万画素でも画質が良くならないことを主張した場合、これに基づく主観評価知識は、知識 (2) と矛盾する。

合理的な意思決定のためには、こうした不完全な知識ベースを適切に扱う方法論が必要となる。

2.3 不完全性を扱う技術

不正確な知識や、断片的な知識の扱いについては、これまでに、数多くの研究が行われている。J. Breese らは、確率値に幅を許した確率区間表現による意思決定の枠組を示した [2]。M. Druzdzel らは、確率区間表現に限らず、確率値の比較表現も含めて、任意の確率値の 2 次確率分布を導出する枠組を示した [5]。P. Cheeseman は、エントロピ最大化の原理に基づいて、断片的な確率値から任意の確率値を確定させる方法を示した [3]。G. Nguetsé らは、断片的な確率値から任意の確率値の上限と下限を計算するアルゴリズムを示した [8]。これらの技術を適用することにより、知識が不正確で断片的であっても、確率値の上限と下限、あるいは、なんらかの基準により確定された確率値に基づいて、合理的な意思決定が可能となる。

矛盾した知識の扱いには、分離に基づくアプローチと、統合に基づくアプローチがある。前者は、知識ベースから無矛盾な知識集合を取り出して利用する (例えば [10, 11])。後者は、喰い違う知識を統合して中間的な知識を生成する (例えば [9, 13])。前者は取り出した知識集合以外は利用しないのに対し、後者は知識ベース中の全ての知識を利用する。したがって、後者は、矛盾していることを意思決定に反映できる点で優れている。一方、後者は獲得された知識を加工して (統合された新しい知識として) 利用するのに対し、前者は獲得された知識を直接利用する。したがって、前者は、ユーザに意思決定の理由を提供するとき、Web 文書中の表現をそのまま利用できるという点で優れている。以下では、それぞれのアプローチについて、適用にあたって重要となる点を述べる。

- 分離アプローチ 無矛盾な知識集合は複数考えられるので、このうちどの集合を利用するかを決める基準が必要となる。これについては、極小や極大の知識集合を利用するという基準が提案されている [1, 11]。しかしながら、これらの基準の妥当性は、対象とした問題の性質に依存しており、常に妥当とは限らない (例えば [7])。利用する知識集合を定めるには、対象とする問題のもつ性質に立ち戻り、これに基づいた基準を採用する必要がある。
- 統合アプローチ 喰い違いのある知識を統合するためのモデルが必要となる。これについては、喰い違いの生じる原因を測定ノイズや解釈誤りとしてモデル化して統合する方法が提案されている [6, 9]。これらの方法は、いずれも、知識を使って 2 次確率分布 (確率値についての確

率分布) を推定し、これにより適切な確率値を導出する。2 次確率分布の推定にあたっては、分布の形や知識の独立性など、適切な仮定を導入する必要がある。

3. おわりに

本稿では、統計的意思決定に必要な確率分布について、これを Web 文書のもつ主観評価知識を使って獲得する方法論を論じた。今後は、本稿で論じた確率分布獲得のモデルを確立するとともに、図 1 に示す意思決定系の各要素のモデル化を進める。そして、意思決定系に関する重要ないくつかの性質 (例えば、信頼できる意思決定に必要とされる知識源や、知識獲得システムの獲得精度など) について解明を進める。

参考文献

[1] Salem Benferhat, Didier Dubois, and Henri Prade. How to infer from inconsistent beliefs without revising? In *Proceedings of the 14th International Joint Conference on Artificial Intelligece (IJCAI-95)*, pp. 1449–1454, 1995.

[2] John S. Breese and Kenneth W. Fertig. Decision making with interval influence diagrams. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pp. 467–478. Elsevier Science Publisher, 1991.

[3] Peter Cheeseman. A method of computing generalized bayesian probability values for expert systems. In *Proceedings of the 6th International Joint Conference on Artificial Intelligece (IJCAI-83)*, pp. 198–202, 1983.

[4] 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*, pp. 509–516, 1998.

[5] 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.

[6] Kazunori Fujimoto, Kazumitsu Matsuzawa, and Hideto Kazawa. An elicitation principle of subjective probabilities from statements on the internet. In *Proceedings of the 3rd International Conference on Knowledge-Based Intelligent Information Engineering Systems (KES-99)*, pp. 459–463, 1999.

[7] Hwee Tou Ng and Raymond J. Mooney. On the role of coherence in abductive explanation. In *Proceedings of the 9th National Conference on Artificial Intelligece (AAAI-90)*, pp. 337–342, 1990.

[8] Guy-Blaise Douanya Nguetśe, Pierre Hansen, and Brigitte Jaumard. Probabilistic satisfiability and decomposition. In C. Froidevaux and J. Kohlas, editors, *Lecture Notes in Artificial Intelligence 946*, pp. 151–161. Springer-Verlag, 1995.

[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] Marco Ramoni and Alberto Riva. Belief maintenance in Bayesian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*, pp. 498–505, Seattle, WA, 1994.

[11] Raymond Reiter and Johan de Kleer. Foundations of assumption-based truth maintenance systems: Preliminary report. In *Proceedings of the 6th National Conference on Artificial Intelligece (AAAI-87)*, pp. 183–188, 1987.

[12] Arnaud Sahuguet and Fabien Azavant. Wysiwyg web wrapper factory (w4f). Available from `http://db.cis.upenn.edu/W4F/`, 1999.

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

[14] 繁桝算男. 意思決定の認知統計学. 朝倉書店, 1995.

[15] 市村, 長谷川, 渡部, 佐藤. テキストマイニング：事例紹介. 人工知能学会誌, Vol. 16, No. 2, pp. 192–200, 2001.

[16] 藤本, 賀沢, 佐藤, 島津, 北. ネット情報を使った意思決定支援 dsiu における知識獲得技術. 人工知能学会論文誌, Vol. 16, No. 1, pp. 120–129, 2001.

[17] 藤本, 山本. ネット情報を使って意思決定する過程の数理モデルの提案. 人工知能学会研究会資料 SIG-FAI/KBS-J, pp. 29–33, 2001.

[18] 富浦, 渡辺他. 特集：ここまできた自然言語処理. 情報処理学会誌, Vol. 41, No. 7, pp. 762–796, 2000.