

学習履歴データベースを用いた 学習支援システム用 Causal Network の自動構築法

大西 佳亮 藤原 祥隆 岡田 信一郎 倉内 誠 吉田 秀樹

北見工業大学

我々は、Causal Network(Bayesian Network と呼ばれる)を用いた確率的推論により学習者の苦手分野を推論し、苦手分野を優先的に学習させる学習支援システムを提案してきた。これまで、このシステムの Causal Network の構築は教材作成者の専門知識に依存して行われてきたが、本研究では実際に学習者が問題を正解できたかどうかを集めた学習履歴データベースを用いて、より客観的な Causal Network を自動的に構築する方法を提案する。

Automatic Constructing Method of Causal Network Using Learning Record Database for Learning Assistance System

Yoshiaki OHNISHI Yoshitaka FUJIWARA Shin-ichiro OKADA
Makoto KURAUCHI Hideki YOSHIDA

Kitami Institute of Technology

We have proposed a learning assistance system with a learning strategy of weak point conquering, in which the target user's weak point is automatically detected using a Causal Network (CN) inference technology. In this system, the constructor of the CN is completely dependent on the knowledge of a human expert of the problem domain.

In this report, an objective method of automatically constructing a CN structure is proposed using an answer record database containing the correct or not-correct results of exercise problems executed by the learners of this system.

1 はじめに

幅広い内容を持つ分野の学習を行うには、得意科目を伸ばすよりも苦手科目を克服する方が学習効果が高いと考えられる。我々は確率的推論技法である Causal Network[1](Bayesian Network とも呼ばれる)を用いユーザの苦手な分野を優先して学習させる学習支援システムを提案してきた [2]。これまで Causal Network の構築は教材作成者の専門知識に依存してきたが、教材作成者にとって負担が大きく、また、実際の学習者の苦手分野の傾向を反映していない可能性がある。そこで本研究では実際に学習者が問題を正解できたかどうかを集めた学習履歴データベースを用いて、より客観的な Causal Network を自動的に構築する方法を提案する。

2 Causal Network

Causal Network (以下 CN) とは、事象をノードとし、事象間の因果関係を表すエッジでそれらを結合して DAG(非循環有効グラフ)を構成し、ある事象について分かった事実から他の事象の確率的状態がどのようになるかを推論する確率的技法である。

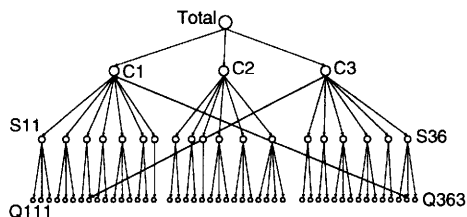


図 1: 一般的な学習支援システムの構造

図 1 に我々が扱う一般的な学習支援システムの CN 構造を示す。この構造では各ノードを Total、Chapter (C1~C3)、Section (S11~S36)、Question (Q111~Q363) と階層的に分類している。Question は問題を表し 正解/

不正解の2つの状態を持ち、それ以外のノードは問題の分野を表し 理解している/理解していないの2つの状態を持つ。Total、Chapter、Section の理解している確率を理解度と呼び、理解度からユーザの苦手な分野を推測している。各問題 (Question) は1つの分野 (Section) に属しているが、問題によっては複数の分野に関連のあるものもあり、その場合は関連のある分野 (Chapter) と問題 (Question) をエッジでつなぐ。

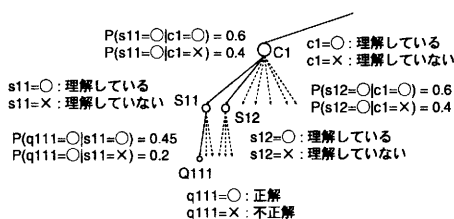


図 2: Causal Network の一部

図 1 の一部分である図 2 を例に取り、CN の推論の例を示す。以下、ノードの値が「正解」又は「理解している」のときは「○」、「不正解」又は「理解していない」のときは「×」で表す。条件付確率 $P(q111=○ | s11=○)$ は S11 が理解されていると分かったとき ($s11=○$) の Q111 が正解になる ($q111=○$) 確率を表している。このとき、「Q111 を正解した」という事実が分かったとする。その事実 $P(q111=○) = 1.0$ を CN に入力すると、CN は与えられた条件付確率を元に、S11 を理解している確率 $P(s11=○)$ 、C1 を理解している確率 $P(c1=○)$ 、S12 を理解している確率 $P(s12=○)$ 等を計算で求めることができる。計算手順は省略するが、「Q111 を正解した」という事実だけが分かった場合の計算結果は $P(s11=○) = 0.77$ 、 $P(c1=○) = 0.5$ 、 $P(s12=○) = 0.6$ となる。

3 Causal Network 自動構築の 手順

今まで、問題が複数の分野に関連する場合に存在する Chapter - Question 間のエッジを発見するために、教材作成者が全ての問題に対して自分が属していない分野との関連を検討していた。しかし、問題数が多い場合、教材作成者の負担が大きいという問題がある。そこで、実際の学習者が問題を正解できたかどうかというデータを集めて学習履歴データベースを作成し、それを元に関連のある分野 (Chapter) と問題 (Question) を推論し、CN を自動的に構築する方法を提案する。

3.1 学習履歴データベース

学習履歴データベースとは複数の学習者に全ての問題を解いてもらい、問題に正解したかどうかを集めたものである。表 1 にデータベースの例を示す。

	学習者			
	A	B	C	...
Q111	○	×	○	...
Q112	×	×	○	...
Q113	○	○	×	...
...				
Q363	○	×	×	...

表 1: 学習履歴データベース

3.2 不足データの補完

関連のある問題と分野の推論方法 (後述) ではデータベースに全てのノードのデータが揃っていないなければならない。しかし、学習履歴データベースには Question ノードのデータしか無いため Total、Chapter、Section ノードのデータを補完する必要がある。以下に Ques-

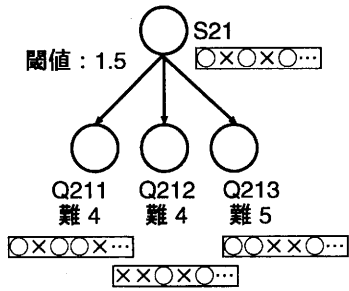
難易度	重み
5	1.4
4	1.2
3	1.0
2	0.8
1	0.6

表 2: 難易度と重みの対応

tion ノードのデータからこれらのデータを補完する方法を説明する。

まず問題の難易度に対して表 2 のような重みを付ける。Section のデータは、各 Section のノードに属する問題のノード数の 2 分の 1 を閾値とし、その Section に属する問題のノードの値が「正解」のノードの重みを足し合わせる。足し合わせた重みの合計が閾値を越えていれば、その Section ノードの値を「理解している」とし、越えていなければ「理解していない」とする。Total と Chapter のデータは、Chapter と Section の難易度を全て 3 として考えることによって、その分野 (Total、Chapter) に属する分野 (Chapter、Section) の値が半分以上「理解している」となっているときにその分野の値も「理解している」となる。このような方法で Section、Chapter、Total の順にデータを補完する。

図 3 のような例を示すと、Section S21 に属している問題は Q211、Q212、Q213 の 3 つなので閾値は $3 / 2 = 1.5$ となる。また、Q211 と Q213 が正解して Q212 が不正解だったとすると、正解した 2 つの問題の難易度は 4 と 5 なので重みは 1.2 と 1.4 となり合計が 2.6 となる。この場合は正解した問題の重みの合計が閾値を越えているので S21 の値は「理解している」になる。



閾値 : 問題数(3) / 2 = 1.5
 データ(先頭) : 1.2 + 1.4 = 2.6
 閾値(1.5) < データ(2.6) より
 S21の先頭のデータは「○」になる。

図 3: データ補完の例

3.3 関連のある分野の推測

学習履歴データベースを使用し関連のある問題 (Question) と分野 (Chapter) を推測する方法として、K2 アルゴリズム [3] に変更を加えたものを使用する。まず K2 アルゴリズムとは、図 4 のように初めにエッジが無くノードだけがある構造があり、全てのノードがとった値をデータとするデータベースを持っている。その状態からデータベースを使用し、関連のあるノードを推測する方法である。

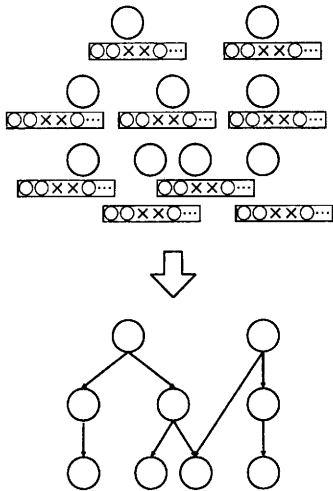


図 4: K2 アルゴリズム

そして、K2 アルゴリズム に以下のような変更を加えたものを使用した。

- 各ノードに深さレベルを持たせ階層構造にする。
- 基準となる構造が存在し、その構造は変化しない。
- 基準となる構造に含まれていない Chapter と Question ノード間の関連だけを推測する。
- 使用するデータベースの Total、Chapter、Section 部分は Question ノードを利用してデータを補完したものになっている。

ここで言う基準となる構造とは、Chapter と Question の間にエッジが無い構造のことである。このような変更を行った K2 アルゴリズム (図 5 参照) を使用し、Question と Chapter の関連を推測することができれば教材作成者の負担を減らすことができる。

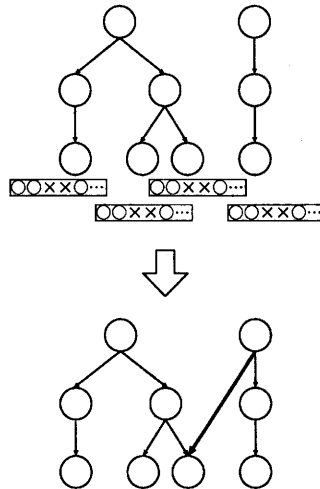


図 5: 変更を加えた K2 アルゴリズム

4 実験

4.1 学習履歴データベースの作成

データベースを生成するには膨大な数のデータが必要になる。今回は実データがないためシミュレーションによって問題の正解/不正解を求め [5]、そのデータを使用し学習履歴データベースを作成した。シミュレーションの方法は、まず推論によって求めたい CN の構造を用意する。実験のため図 6, 7 のような 2 つの構造を用意した。

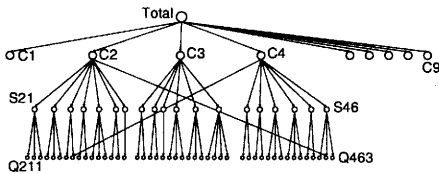


図 6: 推論により求めたい構造 1

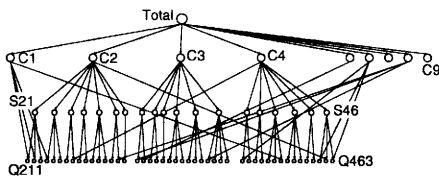


図 7: 推論により求めたい構造 2

これらの構造の関連のあるノード間に条件付確率を設定する。各 Chapter と各 Section の条件付確率は共通の値になっており、Question は問題の難易度によって違う値になっている。この条件付確率を使用し、1 番上の階層である Total から順に下の階層に向かって理解している/理解していないの値と正解/不正解の値を推測していく。

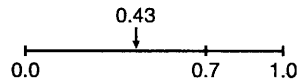
Section S21 を例にとると図 8 のようになる。S21 は上の階層で関連のあるノードは Chapter C2 だけである。S21 の条件付確率は $P(S21 = \bigcirc \mid C2 = \bigcirc) = 0.7$ 、 $P(S21 = \bigcirc \mid C2 = \times) = 0.3$ である。このとき、C2 が「理解してい

$$\begin{aligned} P(S21 = \bigcirc \mid C2 = \bigcirc) &= 0.7 \\ P(S21 = \times \mid C2 = \bigcirc) &= 0.3 \\ P(S21 = \bigcirc \mid C2 = \times) &= 0.3 \\ P(S21 = \times \mid C2 = \times) &= 0.7 \end{aligned}$$

↓ C2 = ○ が確定

$$\begin{aligned} P(S21 = \bigcirc \mid C2 = \bigcirc) &= 0.7 \\ P(S21 = \times \mid C2 = \bigcirc) &= 0.3 \end{aligned}$$

↓ 0.0~1.0の間で
ランダムな値を決定
例：0.43



↓

0.0~0.7の間に入っているので
S21の値は「理解している」になる。

図 8: データの推論法の例

る」と推論できたとすると S21 が「理解している」となる確率は 0.7 となる。そして、0.0 ~ 1.0 の間でランダムな値を決定し、その値が 0.0~0.7 の間であれば S21 の値を「理解している」とし、0.7~1.0 の間であれば S21 の値を「理解していない」とする。

このような方法で全てのノードの値を推測し Question ノードの正解/不正解の値だけを実験用データとして取り出す。これで学習者一人分の正解/不正解パターンが作成され、これを学習者分繰り返し学習履歴データベースを作成した。

4.2 実験結果

図 6 と図 7 の構造からそれぞれ 20 種類のデータベースを 4.1 のシミュレーションによって求め、図 9 の構造を基準となる構造として 3 節で説明した手法を適用した。

そして、ノード間の関連が推測された確率をいくつか抜粋して図 6 の構造の場合は図 10 に、図 7 の構造の場合は図 11 に示す。

図 10 の場合には関連があると仮定したノ

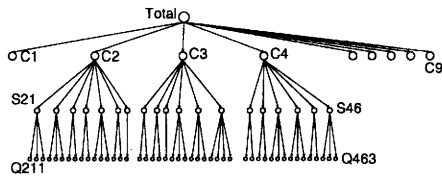


図 9: 基準となる構造

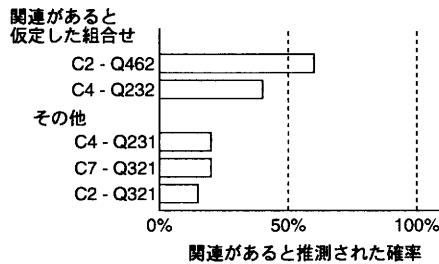


図 10: 実験結果 1

ド(図 6 でエッジでつながっている Question と Chapter のノード)間に関連があると推測された確率は、それ以外のノード間で関連があると推測された確率に対してある程度高い確率を示している。この結果だけを見ると問題と分野の関連を推測できたとと言えるかもしれない。

しかし、図 11 の場合には関連があると仮定したノード(図 7 でエッジでつながっている Question と Chapter のノード)間に関連があると推測された確率の最小値は 30%、それ以外のノード間で関連があると推測された確率の最高値は 70% であり 2 つの結果に明確な境界がない。このような結果になるのは、問題数が多く関連のある問題と分野も多いために難易度が同じ問題だと解答結果が似たようになるものが出てくるためだと考えられる。そのため関連があると仮定した問題と分野を完全に推測できると言うことはできないが、関連があると仮定した問題と分野は全てある程度の確率で推測されているため関連がある問題と分野の組合せの候補は大幅に減少したと言える。

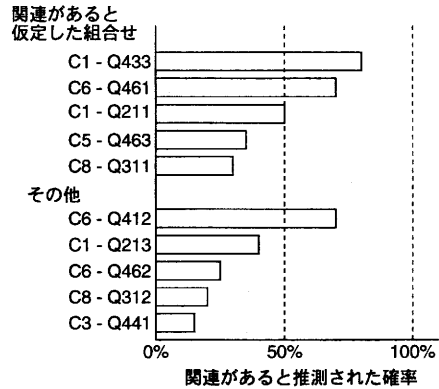


図 11: 実験結果 2

よって実験によって推測された問題と分野の組合せに対してだけ関連があるかどうかの検討を行うことで、教材作成者の負担が減ると考えられる。

5 まとめ

これまで CN の構築は教材作成者の専門知識に依存してきたため、教材作成者の負担が大きい、実際の学習者の苦手分野の傾向を反映していない可能性がある等の問題点があった。この問題点を解消するため学習履歴データベースを用いて、より客観的な CN を自動的に構築する方法を提案した。

今回の実験結果から CN の構築を完全に自動化するのは難しいが、関連がある問題と分野の組合せの候補を大幅に減らすことができ、教材作成者の負担を減らすことが出来ると言える。さらに、問題と分野の関連を検討する回数が減ることによって 1 つの検討に今までよりも時間を掛けることができ、より正確に問題と分野との関連を分析できると考えられる。

参考文献

- [1] Richard E. Neapolitan : “PROBABILISTIC REASONING IN EXPERT SYSTEMS (THEORY AND ALGORITHMS)”, JOHN WILEY & SONS, INC. 1990.
- [2] 藤原 祥隆, 松澤 文太郎, 岡田 信一郎 : “ユーザ適応化機能を備えた学習支援システム”, 情報処理学会研究報告, vol.2001, no.122, pp.17-24, Dec.2001.
- [3] G.F.COOPER and E.HERSKOVITS : “BAYESIAN INDUCTION OF PROBABILISTIC NETWORKS”, Kluwer Academic Publishers(1992).
- [4] 大西 佳亮, 藤原 祥隆, 岡田信一郎, 吉田 秀樹 : “不完全データベースを使用した学習支援システム向け Causal Network の自動構築法”, 情報科学技術フォーラム 2003, 2003.
- [5] Fung R., Chang K. : “Weighing ang Integrating Evidence for Stochastic Simulation in Baysian Networks”, Uncertainty in Artificial Intelligence 5, pp209-219, 1990.