階層構造を持つ属性の組とクラスで与えられる構造規則のクラス分類

中井 慎司　池原 悟　村上 仁一
鳥取大学大学院工学研究科
{nakai,ikehara,murakami}@ike.tottori-u.ac.jp

1 はじめに

自然言語処理分野では、意味に関する知識を表現するためにis-a関係を使用し、木構造に表現する。名詞を例に挙げると、名詞間の関係は、is-a関係およびhas-a関係を用いシソーラスの形で表現される。このシソーラスは、種々の意味解析に使用されている。一例として名詞句「AのB」の意味関係の付与について考えてみる。標本から得られた例が「彼の自転車：所有関係」、「先生の机：所有関係」の2つ場合、この用例をそのまま辞書に登録するのではなく、「彼」「先生」を＜人＞、「自転車」「机」を＜人工物＞に汎化し、「＜人＞の＜人工物＞：所有関係」を辞書に登録した方がよいだろう。

このような構造規則を人手で作成するのはコストや時間などの問題で困難であることを考慮に入れ、学習によって自動的に構造規則を獲得する手法を開発する必要がある。

従来、木構造属性を許容する決定木学習 [1] により、階層構造を持つ属性を扱えるようになったが、ルールがツリー構造で出力されるため、人間がみても理解しづらい。また、従来は包含関係を持つ構造規則を生成することが困難であった。

そこで、本論文では、階層構造を持つ属性の組とクラスで与えられる構造規則のクラス分類を行うための構造規則をボトムアップにより自動生成する学習アルゴリズムを提案する。本手法は、以下の2点の条件を満たす。

1. 構造規則に記述されている属性の汎化を行い、より一般的な規則を生成する
2. クラス分類に有効な属性のみを選択し、より少ない属性の組で規則を記述する

また、包含関係を持つ構造規則が生成できるように拡張する方法を提案する。これにより例外規則から順次汎用規則まで生成できる。

本論文の構成は以下の通りである。2章でボトムアップによる構造規則の自動生成法について述べ、3章で包含関係を持つ構造規則が生成されるように拡張する。4章では木構造属性を許容する決定木学習との比較を行い、最後に5章で本論文をまとめる。

2 ボトムアップによる構造規則の生成

ボトムアップによる構造規則の生成を考えるとき、以下の2つの条件を満たすように規則の生成を行う。

1. 構造規則に記述されている属性の汎化を行い、より一般的な規則を生成する
2. クラス分類に有効な属性のみを選択し、より少ない属性の組で規則を記述する

次節より、上記の2点について具体的にアルゴリズムを説明していく。

2.1 汎化を用いた構造規則の生成

汎化を用いた構造規則の生成の手順は、まずすべての学習データを最終的に得られる構造規則と仮定し、それらの構造規則を少しづつ汎化して一般的な構造規則を生成していく。ここで汎化とは、階層構造を持つ属性のあるノードをひとつ上の親ノードに変更することにより、より抽象度の高いノードにしていく操作である。

汎化の過程で、汎化の対象としている属性が以下の2つの条件のどちらかを満たした時、その属性に対する汎化を終了させる。

そして、最終的にどの属性もそれ以上汎化できなくなった時、終了させる。

1. あるノードを汎化した時、他のクラスの領域と重複した場合（図1参照）

2. あるノードを汎化した時、その汎化によって新しい用例が包含されない場合

1. の条件については、実際にはノイズを考慮に入れ、次のように条件を緩くするのが良いと思われる。

あるノードを汎化した時、他のクラスの領域を包含する場合でも、不正解率がある一定の閾値内であれば、他のクラスの用例をノイズとみなし、汎化していく。

これは決定木の枝刈り（pruning）に相当する。この時の最適な閾値は、学習データに依存するので実験的に得る必要がある。

2. の条件は、構造規則の記述は、より広い範囲をより詳細な規則で記述するという一般的な規則の記述の方針に従って、それ以上汎化可能であっても、汎化をストップする。これにより、必要な範囲だけをカバーする構造規則が得られ、人間がみて理解しやすい構造規則が得られる。なお、テストデータに構造規則を適用する時、どの構造規則にも当てはまらない場合もあるので、デフォルトルールを作成しておく必要がある。

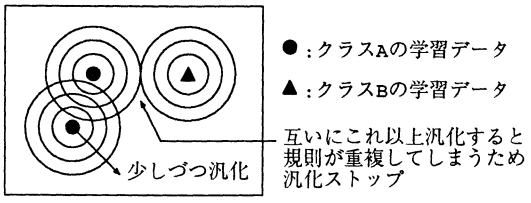

図 1: 汎化の過程の例

2.2 クラス分類に有効な属性の選択

2.2.1 複数の学習データの作成

クラス分類に有効な属性の選択を行うために、あらかじめ属性の組み合わせによって複数の学習データを作成する。

この時、n 個の属性からなる学習データでは $(2^n - 1)$ 個の学習データが作成される。

例として、以下のような 3 つの属性および 1 つのクラスからなる学習データを考える。

(X,Y,Z:C)
X,Y,Z:属性,C:クラス

3 つの属性の組み合わせによって以下の 7 つの学習データを作成する。
(X:C),(Y:C),(Z:C)
(X,Y:C),(Y,Z:C),(X,Z:C),(X,Y,X:C)

得られたぞれぞれの学習データに対し、2.1 節のアルゴリズムに従い、汎化を用いた構造規則の生成を行う。

2.2.2 クラス分類

本節では、テストデータのクラス分類を考える。この際、2.2.1 節で複数の構造規則が生成されるため、あるテストデータに対し、複数の構造規則が適用される可能性がある。複数の構造規則が適用された場合でも、そのクラスが同一の場合は問題ないが、クラスが異なる場合、何らかの規則によってクラスを一意に決定する必要がある。

この場合の対処の仕方には次の 2 つが考えられる。

1. 複数適用された場合は、多数決をとり一番多いクラスをテストデータのクラスに決定する方法
2. 構造規則に信頼度 (未知の用例に対する予測正解率) を付与しておき、一番信頼度の高い構造規則のクラスをテストデータのクラスに決定する方法（Yarowsky の決定リスト [2] に相当する）

3 包含関係を持つ構造規則の生成

この章では、汎化を用いた構造規則の生成法について包含関係を持つ構造規則が生成できるように拡張する。

3.1 包含関係を持つ構造規則について

包含関係を持つ構造規則の生成は階層構造を持つ属性を使用するため、汎化によって得られる構造規則を包含関係を持つように拡張することは可能である。

包含関係を持つ構造規則を以下の 2 つに分類する。

1. 上位下位関係が成り立つ場合の包含関係を持つ構造規則
2. 上位下位関係が成り立たない場合の包含関係を持つ構造規則

1. について図 2 の左図を用いて説明する。

構造規則に包含関係を認めると、以下の 2 つの構造規則の生成だけですむ。
1.（ノード d → B）

2.（ノードa→A）
注：(L→R)：条件Lが成り立つとき、結論Rである。この時、結論Rはクラスである
注：波線のノードの構造規則は生成されない。

これらの規則は、ある属性がノードdに包含されるならばクラスはBである。それ以外でノードaに包含されるならばクラスはAであると解釈する。包含関係を持つ構造規則を生成することにより、規則数を削減することができる。なお、包含関係を持つ構造規則の場合、適用順序を考慮に入れる必要がある。

なお、構造規則に包含関係を認めない場合、以下の3つの構造規則が生成される。
（ノードb→A）
（ノードc→A）
（ノードd→B）

2.について図2の右図を用いて説明する。
（ノードB→A）
（ノードC→A）
上記の2つの規則は（ノードA→A）に包含され省略可能だが、ノードAとノードB,C間に上位下位関係が成り立たない場合は別々の構造規則として出力する。

この例は、汎用規則に包含され、通常ならば出力されない構造規則も出力される例である。これは例外的な例で、日本語語彙体系 [3] において出てくる問題である。中井ら [4] は上位下位シソーラスである日本語語彙体系シソーラスの中で抽象度の高いノード間では単語単位で見て上位下位関係が成り立たないノード対があり、上位のノードが下位のノードに影響を及ぼさないように上位のノードを使用した構造規則を特別規則として、別扱いにする必要があると報告している。よって、上位ノードと下位ノードを使用した構造規則は別の構造規則と考え、たとえクラスが同一で下位ノードを省略できたとしても別の構造規則として出力する。

3.2 包含関係を持つ構造規則の生成法

包含関係を持つ構造規則は次の手順で生成される。

1. 全学習データを初期の構造規則として汎化を開始
2. 汎化を進めていく過程で他のクラスの構造規則を包含する場合、包含される構造規則を出力し、包含される規則に当てはまる学習データを除去し汎化を進める。テストデータに適用する場合には構造規則が出力された順（例外規則から）に適用する。

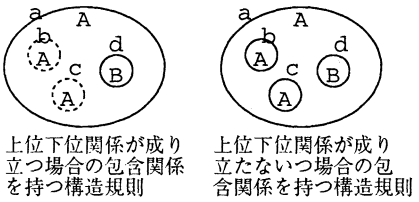

図 2: 包含関係を持つ構造規則

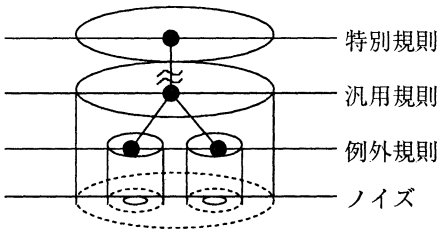

図 3: 包含関係を持つ構造規則の関係

図3に包含関係を持つ構造規則の関係を示す。上記の手順で、基本的には何層（シソーラスの総段数まで）にも包含関係を持つ構造規則が生成可能だが、人間の理解しやすさを考慮に入れれば、例外規則と汎用規則の2層が最適ではないかと思われる。一番下の層は、その構造規則が捨てられるという点で他の層とは異なっている。この層をノイズの層と呼ぶ。ノイズとして捨てられるか、その一つ上の例外規則となるかは、構造規則の抽象度の度合、または構造規則が包含する用例の数に依存する。ある尺度において閾値以下だとノイズと判定され、捨てられ、閾値以上だと構造規則として採用される。

4 本手法と木構造属性を許容する決定木学習との特徴の違い

この章では、決定木学習 [5] において木構造属性を扱えるようにした手法 [1] と本手法との特徴の違いを挙げ、それぞれの長所、短所、またそれぞれ有利な問題領域について考察する。

1. 決定木はトップダウン、汎化を用いた本手法はボトムアップで規則を生成

   この方式の違いによる分類精度はほぼ差がなく、同等の性能が出ると言ってよい。

2. 本手法は学習データを丁度カバーする程度に汎化を行い構造規則を生成（どの規則にも当てはまらない場合はデフォルトルールが適用される）、決定木はすべての空間をカバーするようにツリーを生成

   本手法によって得られた構造規則は、必要以上に汎用的な構造規則を生成せず、人間がみた時、理解しやすい形になっており、生成された構造規則をあとで人手で修正し、構造規則のデータベース化を行う際、容易に行えると考えられる。

3. 学習結果の安定化手法として、決定木でのBoostingに対し、本手法では複数の構造規則を生成し、多数決をとる手法が対応

   春野ら[6]は日本語係り受け解析に決定木を用いている。この中で、Boostingを行えば、さらに解析精度が向上すると報告している。Boostingとは学習データが少し変わっただけで学習結果が大きく変化するという不安定性を解消する手法である。本手法では、2.2.1節で述べているように複数の構造規則を生成し、得られた複数の構造規則の多数決によりクラスを決定する。この手法は学習結果の不安定を解消するという点で決定木におけるBoostingに対応すると考えることができる。

4. 包含関係を考慮に入れた構造規則は自然言語処理の諸問題に対し有効

   ここでは、自然言語処理の諸問題のなかでも、日本語名詞句「AのBのC」の係り受け解析[7]を考える。この名詞句の係り先を決定する構造規則を生成した場合、次のような特徴がある。それは、まずいくつかの例外規則があり、その例外規則に当てはまるテストデータを除いたあとに、それらを包含する汎用規則が適用されるという特徴である。

   包含関係を考慮に入れた構造規則の生成は、3章に示した手法に従って、明示的に行われるため名詞句などにおける構造規則の生成には有効と考えられる。なお木構造属性を扱える決定木でも包含関係を表現できるが明示的に行われないため、実際には包含関係を含んだツリーは生成されにくい。

5 おわりに

本論文では、階層構造を持つ属性の組とクラスで与えられる構造規則のクラス分類を行うための構造規則の学習アルゴリズムを提案した。本手法は、1)構造規則に記述されている属性の汎化を行い、より一般的な規則を生成する、2)クラス分類に有効な属性のみを選択し、より少ない属性の組で規則を記述する、の2点の条件を満たす。また、包含関係を持つ構造規則が生成できるように拡張した。これにより例外規則から順次汎用規則まで生成できる。本手法を利用することにより、クラス分類に使用する属性に階層構造を持つ属性の例として名詞シソーラスの意味属性を使用することが可能になり、名詞句解析など、複数の語から構成される表現の構造規則の自動生成に役立つと期待される。

参考文献

[1] フセイン・アルモアリム, 秋葉泰弘, 金田重郎：木構造属性を許容する決定木学習, 人工知能学会誌, Vol. 12, No. 3, pp. 421–429 (1997).

[2] Yarowsky, D.: Decision lists for lexical ambiguity resolution, *32th Annual Meeting of the Association for Computational Linguistics*, pp. 88–95 (1994).

[3] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙体系, 岩波書店 (1997).

[4] 中井慎司, 池原悟, 白井諭：「の」型名詞句における品詞情報と意味情報を併用した係り受け規則の自動生成, 情処研報, Vol. 98-128-7, pp. 45–51 (1998).

[5] Quinlan, J.: AIによるデータ解析, トッパン (1995).

[6] Haruno, M., Shirai, S. and Ooyama, Y.: Using Decision Trees to Construct a Practical Parser, *COLING-ACL'98*, Vol. 1, pp. 505–511 (1998).

[7] 中井慎司, 池原悟, 白井諭：「の」型名詞句における名詞間の係り受け規則の自動生成法, 信学技報, Vol. 98, No. 53, pp. 15–22 (1998).