平成 12 年度卒業研究

入学年度　平成 9 年度	学籍番号　09117963	氏名　山田 泰大
論文題目　事例に基づく学習における背景知識の利用に関する研究		犬塚 研究室

1　はじめに

事例に基づく学習 (IBL) は類似した事例の属するクラスを参考に分類を行なう。その為、事例間の類似度 (距離) 尺度が必要となる。事例の一階述語表現は柔軟な表現方法であるが、その柔軟さ故に距離尺度の求め方が問題となる。近年、一階表現された事例間の距離を帰納する DISTILL[1] というシステムが提案されている。一階述語論理を用いた学習手法 (ILP) では背景知識を利用できる事が重要な利点である。DISTILL ではこの点で十分とは言えない。本研究は IBL において背景知識を利用する方法を提案する。

2　仮説に基づく距離尺度の帰納 (HDD)

距離尺度の帰納　(1)$class(E_i) \neq class(F_i)$ なる事例 E_i,F_i を訓練事例からランダムに選ぶ (2)E_i をカバーし、F_i をカバーしない最も一般的な仮説 h_i(事例 E_i の特徴とみなせる) を求める (3)(1),(2) を d 回繰り返し、仮説集合 $\mathcal{H}=\{h_1,\cdots,h_d\}$ を作成する (これを事例間の距離を求める際の尺度とする)

事例間の距離の求め方　事例 x_q を d 次元ベクトルへ次のように写像する (1) 事例 x_q が仮説 h_i をどれだけ満たしたかの評価値をベクトルの i 番目の成分の値とする (2)(1) を $i=1\sim d$ に対して行なう事により事例を $\mathcal{N}^d$ のベクトル空間へ写像する (3) 事例間の距離はこのベクトル間のユークリッド距離として求められる

HDD の特徴は、事例の記述全体からベクトルへの写像を決定するため、よく使われる重み和による距離の求め方と違い局所的な事例の記述のみに距離が依存しない事である。実際問題ではこの様な距離尺度が求められている。

3　背景知識の利用に関する 3 つの提案手法

提案手法 1:飽和節による知識の埋め込み

事例の飽和節とは、背景知識により得られる全ての知識を事例の記述に埋め込んだものである。事例の飽和節を求める事により、あらかじめ事例に背景知識を埋め込むことを提案する。このことによって、DISTILL により良い距離尺度を帰納させることを目指す。

提案手法 2:背景知識を利用した距離の補正

事例が類似しているほど、2 つの事例は共通のルールを満たすという仮定に基づき DISTILL により求められた距離を式 (1) で補正する。ここでは x,y は事例、R はルールの集合,cover(x,R) は式 (2) で定義する。

$$dist(x,y) \stackrel{\text{def}}{=} dist_{\text{DISTILL}}(x,y)\cdot\frac{|\text{cover}(x,R)\bigcup\text{cover}(y,R)|}{|\text{cover}(x,R)\bigcap\text{cover}(y,R)|} \quad (1)$$

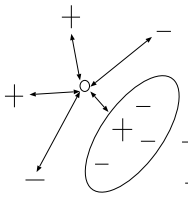
$$\text{cover}(x,R) \stackrel{\text{def}}{=} \{r\in R|r\text{が}x\text{をカバーする}\} \quad \cdots \quad (2)$$

提案手法 3:事例とルール間の距離の定義とそれを用いた分類

X を事例空間とする。ルール r と事例 x の距離を式 (3) で定義する。

$$dist(x,r) \stackrel{\text{def}}{=} \min_{e\in X_r} dist_{\text{DISTILL}}(x,e) \quad \cdots \quad (3)$$

$$X_r \stackrel{\text{def}}{=} \{x\in X|r\text{が}x\text{をカバーする}\} \quad \cdots \quad (4)$$



benzene(X):-atom(X,A,c,E),..

T を未知の事例集合とする。事例 $x_q\in T$ と、ルール r によりカバーされる事例のうち最も近い事例 $e\in T_r-x_q$ との距離を事例 x_q とルール r の距離とする。そして、このルールにカバーされる訓練事例のクラスの割合に応じルールも投票に参加する。

図 1: 事例とルールの距離

4　実験

化学物質の突然変異誘発性物質判定問題 [2] で提案手法 1〜3 を評価 (10-fold cross-validation) する。表 1に突然変異誘発性問題における分類精度を示す。化合物は複雑な構造をしているため属性値表現が困難であり C4.5 などの属性に基づく学習法では扱い難い。

表 1: 各提案手法の実験結果

	精度		
	ルール 1	ルール 2	ルール 3
提案手法 1	89.9%	89.9%	93.0%
提案手法 2	91.0%	87.8%	
提案手法 3	92.6%	92.0%	

パラメータ $d=70$,$\eta=300$,$K=30$

ルール 1 あるリング構造があるか、ないかだけの知識
ルール 2 リング構造の数を教える知識に対するルール
ルール 3 同じリングが複数存在するとき、そのリングがある旨を繰り返し返す知識

5　まとめ

本研究では関係を用いた知識を与えることができるようになった。その知識が部分的ではあっても精度の高い知識であれば、事例に基づく学習とその知識が協調することによりより良い精度が得ることができるであろう。

参考文献

[1] Michéle Sebag. *"Distance Induction in First Order Logic"*, ILP,pp.264-272,1997.

[2] http://oldwww.comlab.ox.ac.uk/oucl/groups/machlearn/mutagenesis.html