遺伝的概念獲得におけるノイズの対応に関する考察

6N−7

遠藤　聡志　　大内　東

北海道大学工学部情報工学科

1　はじめに

経験的な事柄から過去の出来事を説明したり未来を予測するような，ある一般性を導くような推論は帰納推論 (Induction) と呼ばれる．この帰納推論の典型例として Mitchell 等によって提案された概念形成問題がある [Mitchell 77]．本稿では，この概念形成問題に遺伝的アルゴリズム [Goldberg 89][Holland 75] を応用した遺伝的概念獲得が，従来の概念形成アルゴリズムでは困難とされていたデータノイズへの対応が比較的容易になされることを示す．

2　概念形成問題とデータノイズ

概念形成問題は，帰納推論の典型例の一つであり，以下の形で定式化される．

$$CFP < P, N, C, \Lambda >$$

ここで，Pは概念の正の例，Nは概念の負の例，Cは概念バイアス，Λは論理バイアスである．Version Sapce 法として知られる Mitchell のアルゴリズムは，Λを単一連言形式とし，Cを用いてΛによって記述されるすべての概念を Version Space で表現した概念形成法である．

概念形成のために用いられる事例 P, Nは，例えば教師や自然といった外部から与えられる．概念形成問題において，これらのデータは完全に正しいことが前提であり，Version Space では誤りを含む事例を与えたときには，その誤りによっては仮説を導くことが出来ない場合もある．誤りの無いデータの提示や例外的な観測の排除が困難な場合には，そのようなデータから”もっともらしい”仮説を導く柔軟な概念形成法が必要となる．

Noise Reduction of Concept Learning using GA
Satoshi ENDOH and Azuma OHUCHI
Faculty of Engineering, Hokkaido University

3　遺伝的概念獲得

3.1　単一連言概念獲得アルゴリズム

Mitchell の提案した VersionSpace は獲得すべき概念を単一連言に限定した概念獲得法である．これに対応する GA を用いた概念獲得のアルゴリズムは以下のようになる．

単一連言概念獲得アルゴリズム

Step1: 環境 ENV=<P,Q>の設定;//P=正例集合;Q=負例集合

Step2: 世代数 $t = 0$; 正例を用いた初期世代集団 I=$\{i_1, i_2, \cdots, i_{size}\}$ の生成;

Step3: 各個体の適応度評価;

Step4: 適応度を基準とした 2 個体の選択;

Step5: 2 個体に対する交叉; 確立 r での突然変異;

Step6: 致死判定; 及び集団の入れ替え;

Step7: t++;if(!((適応度=最大値)or(t=最大 Loop 回数))) step3 へ;

Step8: 適応度=最大値となる個体を獲得;

Step9: 獲得概念表示;

アルゴリズムの Step3 において適応度関数は以下の方針の基に設定される．

(a) 適応度関数は多くの正例を説明する個体に高い評価を与える

(b) 適応度関数は，一つでも負の事例を説明する個体に最も低い評価を与える

$$Fitness(pos, nega) = pos - \alpha \cdot nega$$

$$\Downarrow (\alpha \gg 1)$$

$$Fitness(pos) = \frac{pos}{poss_{all}} \ or \ dead$$

3.2　選言概念獲得への拡張

VersionSpace 及び GA による概念獲得に共通の問題点として，その獲得概念記述能力の弱さがあげられる．この問題点に対して，単一連言概念獲得アルゴリズムは基本的に適応度関数を変更することで論理バイアスΛを選言記述に拡張することが出来る．

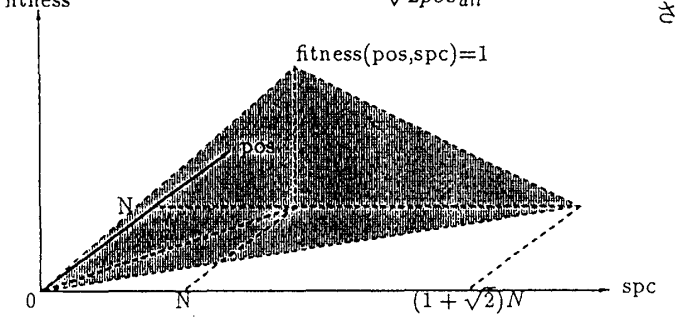
(a) 適応度関数は，多くの正事例を説明する個体に高い評価を与える

(b) 適応度関数は，負事例を一つでも説明する個体には最も低い評価を与える

(c) 適応度関数は，個体が表現する空間を正例が占める割合が高い個体に高い評価を与える．

上記の条件を満足するような適応度関数を以下に示す．

$$Fitness(pos, spc) = max(0, \frac{(1+\sqrt{2})pos - spc}{\sqrt{2}pos_{all}})$$



この関数は，ある個体 (概念記述) が与えられた正例をすべて説明し，さらにその個体が説明しうる事例がないようなもの (正例密度=1) に最も高い評価 (Fitness=1) を与えている．また，説明可能な正例の増加および正例密度の減少に関しては，それぞれ単調増加および単調減少となるように評価すればよいが，この適応度関数では，説明する正例の一増加分と正負未判定の事例の一減少分を等価に評価するように設定されている．

拡張アルゴリズムを以下に示す．

選言概念獲得アルゴリズム

Step1: 環境 ENV=<P;Q>の設定; 獲得概念 C=ϕ;//P=正事例集合;Q=負事例集合

Step2: 世代数 t = 0; 正例を用いた初期世代集団 I=$\{i_1; i_2; \cdots; i_{size}\}$ の生成;

Step3: 各個体の適応度評価;

Step4: 適応度を基準とした 2 個体の選択;

Step5: 2 個体に対する交叉; 確立 r での突然変異;

Step6: 致死判定; 及び集団の入れ替え;

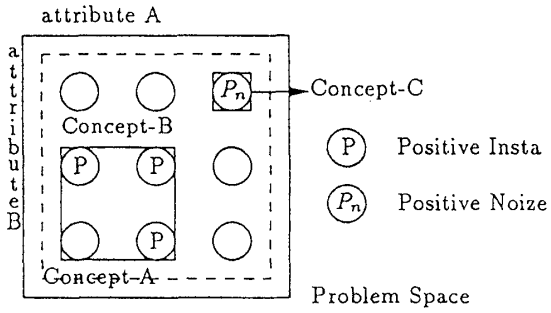
Step7: t++;if(t!=最大 Loop 回数 T) step3 へ;

Step8: 最大適応度を示す i_xを獲得概念に追加;//C=C$\bigwedge\{i_x\}$;
i_xが含む正事例を P から削除;if(P!=ϕ) Step2 へ;

Step9: 獲得概念表示;

3.3 選言概念獲得アルゴリズムのノイズへの対応

選言概念獲得アルゴリズムの適応度評価関数では，個体が表現する空間中に占める正例の割合，すなわち正例密度の概念を導入して各個体を評価している．この結果，例えば図のような事例空間の上では，仮説 A は，0.02 の評価値，仮説 Bは 0.43，仮説 Cは 0.25 となり，通常仮説 Bが連言集合に獲得される．その後，仮説 Cの評価値は 1.00 となり連言集合に加えられる．よって，この適応度関数によって獲得される概念は A ではなく $B \vee C$である．ここで，Cは正例のノイズデータであり獲得概念中の単一連言の一つとして抽出される．



このように選言概念獲得では正例密度の導入が過度の一般化をさけるため，ノイズデータを含めた連言記述に比べ，正しいデータを記述する連言とノイズデータの連言の選言として概念が獲得される可能性が高い．この結果は，利用者のノイズデータ発見に役立つ．

4 おわりに

遺伝的アルゴリズムによる概念形成が，事例中に含まれるノイズに対して柔軟に対応できることを示した．本研究は文部省科学研究費 (奨励研究 (A)No.05750376) の補助を受けている．

参考文献

[Mitchell 77] *Mitchell,T.M. "Version Spaces:A Candidate Elimination Approach to Rule Learning" Proc. of the Fifth IJCAI, 1977*

[Goldberg 89] *Goldberg,D.E. "Genetic Algorithms in Search,Optimization and Machine Learning" Addison-Wesley, 1989*

[Holland 75] *Holland,J.H "Adaptation in natural and artificial systems" Ann Arbor:The Unversity of Michigan press, 1975*