帰納論理プログラミングの繰り返し学習による効率化

6P－4

黒田 洋介　大原 剛三　馬場口 登　北橋 忠宏

大阪大学 産業科学研究所

1 はじめに

近年，帰納論理プログラミング (Inductive Logic Programming: **ILP**) による知識獲得が盛んに研究されている [1]．ILP では，目標概念が与えられた時に，正例，負例，及び背景知識から，全ての正例を導き，1つも負例を導かない目標概念を説明する概念記述を生成することを目的とする．ILP による知識獲得は，表現言語として用いている一階述語論理の表現力の高さから複雑な概念を学習できる反面，大量データからの知識獲得を考えた場合，学習に膨大な時間がかかる傾向がある．そこで，本研究では少量のサンプリングデータを繰り返し学習することで，分類精度を低下させることなく学習時間の軽減を計る手法を提案する．

2 繰り返し学習による学習時間の軽減

大量データから効率的に学習する手法の一つとして，少数の事例をサンプリングして学習することにより学習時間を軽減する手法がある [2]．しかし単純にデータベースから少数の事例をサンプリングする場合，獲得される仮説の分類精度が，全事例を用いて学習する場合に得られる仮説の分類精度と比較して著しく低下する恐れがある．

そこで本研究では，Boosting[4] 手法の考えに基づいて少量の事例を複数回学習することで分類精度の低下を防ぐ手法を提案する．Boosting は 1/2 よりは少し良い確率で分類ができる仮説を学習できる弱学習アルゴリズムを繰り返し用いて学習することにより，高い分類精度を持つ仮説を学習できる強学習アルゴリズムを実現することを目的として提案された手法である．それに対し本研究では，ILP による学習に対して少量のサンプリングデータで学習することにより学習を効率化することを目的とし，分類精度の低下を防ぐ手段として Boosting 手法の考えを利用する．

3 提案手法の概要

本章では，代表的な Boosting 手法である AdaBoost[4] に基づいて ILP の繰り返し学習による効率化を実現する．まず，提案手法の全体の概略図を図 1に示す．提案手法における t 回目の学習では，与えられた事例の確率分布 D_t に基づきサンプリングした少数の事例を ILP を

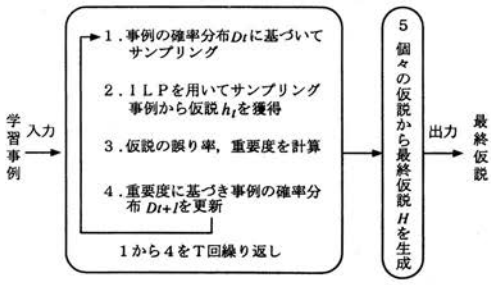

図 1: 提案手法の概念図

用いて学習し，AdaBoost の手法に従って D_t を D_{t+1} に更新する．これを T回繰り返した後，各回で得られた仮説を用いて生成した最終仮説を出力する．以下では，確率分布の更新，及び最終仮説について詳細に述べる．

3.1 誤り率の算出

繰り返し回数 t 回目における事例に対する確率分布を D_t，獲得した仮説を h_t，また事例 i に対する生起確率を $D_t(i)$，i の正負を y_i，仮説 h_tによる i の分類結果を $h_t(i)$ とする．尚，y_i，$h_t(i)$ に関しては i が正例の場合は 1，負例の場合は−1 とし，h_tによって i が正例か負例かを判定できなかった場合 (Unknown の場合)，$h_t(i)$ は 0 を返すものとする．この時，確率分布の更新，及び最終仮説に必要なパラメータとして以下のような誤り率ϵ_t，及び重要度α_tを定義する．

$$\text{誤り率}: \epsilon_t = \sum_{i:h_t(i)=-y_i} D_t(i)$$

$$\text{重要度}: \alpha_t = (1/2)\log((1-\epsilon_t)/\epsilon_t)$$

ここでϵ_tは，誤って分類した事例の生起確率の合計であり，α_tはϵ_tを用いて計算した仮説 h_tの重要度である．もし，$\epsilon_t < 1/2$ であれば$\alpha_t > 0$ であり，α_tはϵ_tが小さい程大きくなり，特に$\epsilon_t = 0$ の時は$\alpha_t = \infty$ が与えられる．

3.2 確率分布の更新

事例に対する確率分布のうち D_1は一様分布として与え，繰り返し回数 t 回目の確率分布 D_tは前節で定義した仮説の重要度α_tを用いて以下のように更新する．

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{\alpha_t} & (h_t(i) = -y_i \text{の場合}) \\ e^{-\alpha_t} & (h_t(i) \neq -y_i \text{の場合}) \end{cases}$$

ここで Z_tは D_{t+1}が確率分布となるように正規化する因子である．このように誤って分類した事例の生起確率を増加させ，正しく分類した事例の生起確率を減少させ

Repeating Learning Process for Efficient Inductive Logic Programming
Yosuke KURODA, Kouzou OHARA, Noboru BABAGUCHI, and Tadahiro KITAHASHI
I. S. I. R., Osaka University

表 1: 実験結果

	実験 1	実験 2
総学習時間 (sec)	967.5	604.025
事例学習時間 (sec)	967.5	531.25
適合率 (%)	98.9386	99.6025
再現率 (%)	99.8478	99.6028

る．これにより，次にサンプリングする際に誤って分類した事例をサンプリングされやすくし，そのような事例を集中的に学習させることで分類精度を向上させる．

3.3 最終仮説の生成

最終的に，繰り返し回数 T回の学習の各回で得られた仮説の集合として最終仮説 Hを生成し，Hによる事例 i の分類結果 $H(i)$ を次のように定義する．

$$H(i) = \begin{cases} positive & (\sum_{t=1}^{T} \alpha_t h_t(i) > 0 \text{ の場合}) \\ unknown & (\sum_{t=1}^{T} \alpha_t h_t(i) = 0 \text{ の場合}) \\ negative & (\sum_{t=1}^{T} \alpha_t h_t(i) < 0 \text{ の場合}) \end{cases}$$

以上のように，最終仮説は与えられた事例に対して，個々の仮説 h_tにα_tを重みとして用いた重みつきの多数決により正負を判定する．この結果，分類精度の低い個々の仮説の集合体である Hを高い分類精度を持つ仮説として扱うことが可能となる．

4 評価実験

提案手法を我々が提案した ILP システム G-REX[3] 上に実装し，評価実験を行った．

4.1 実験内容

本実験においては，実験システムを SUN のワークステーション SS20(hyperSPARC125MHz, 352MB) 上に実装し，UCI で公開されているマッシュルームデータベースを実験対象として用いた．マッシュルームデータベースとは，キノコ 8,124 種について，毒の有無を 22 の属性で記述したデータベースである．本実験では目標概念を「毒を持つ」($poison(A)$) とし，その場合の正例数は 3,916 個，負例数は 4,208 個となる．このデータベースに対して，ランダムに抽出した各 1,000 個の正・負事例を学習セット，残りの事例をテストセットとし，提案手法の有効性を示す為の比較に，学習セットに対して全事例を学習する実験 1 と，繰り返し学習を用いて学習する実験 2 を行った．尚，実験 2 では，学習セットからのサンプリング数を 100 個，学習回数を 10 回とした．また，学習で得られる最終仮説 Hの精度を評価するために，次のような再現率 $P(H)$，適合率 $R(H)$ を求めた．

$$P(H) = \frac{\text{正分類数}}{\text{全事例数}}, \quad R(H) = \frac{\text{正分類数}}{\text{分類数}}$$

ここで，分類数とはテストセット中の事例に対して，仮説が分類することができた事例数，正分類数とはそのうち正しく分類された事例数である．

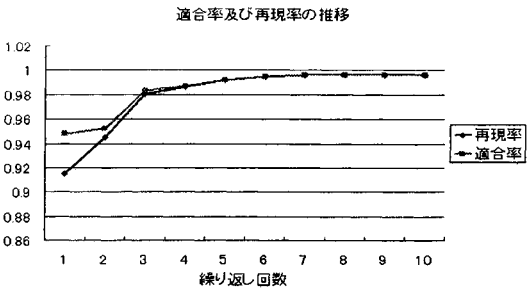

図 2: 繰り返し学習における適合率と再現率の推移

4.2 結果と考察

実験 1,2 を 4 回試行した結果の平均を表 1に，また実験 2 における繰り返し学習の回数と分類精度の関係を図 2に示す．表 1における事例学習時間とは，事例の学習に要した時間であり，総学習時間とは確率分布の更新処理等全てを合わせた時間である．表 1から実験 2 では，実験 1 とほぼ同程度の精度を獲得しながら，実験 1 よりも事例学習時間が 45%が削減され，総学習時間においても 38%削減されていることが分かる．また図 2から，学習を 3，4 回繰り返すうちに分類精度が急激に向上していることが分かる．一方，実験 1 で得られた仮説数が平均 23 個であったのに対して，実験 2 では仮説数が平均 78 個に増えており，学習結果の直感的な分かりやすさが低下する結果となった．

5 まとめ

本稿では，ILP における繰り返し学習を用いた効率化の手法を提案した．本手法では少量の事例をサンプリングして繰り返し学習する際に，少ない学習時間で全事例を用いて学習する場合と同定度の分類精度を持つ仮説の獲得をすることを，特殊な事例を重点的に学習させることで実現し，実験によってその有効性を示した．提案手法は大量データからの知識獲得時など，事例数が膨大になる場合において有効であると考えられる．

今後の課題として，本手法の統計的評価，サンプリング手法や最終仮説の生成手法のさらなる検討等が挙げられる．

参考文献

[1] L. de Raedt : Advances in Inductive Logic Programming, IOS Press, 1996.

[2] F.Provost, D.Jensen, and T.Oates : Efficient Progressive Sampling, In Proceedings of the Fifth International Conference on Knowledge Discovery & Data Mining,pp.23-32,1999.

[3] 高，大原，馬場口，北橋：例外関係に着目した不完全知識の獲得システムの実装と実験的評価, 人工知能学会研究報告 (SIG-FAI-9803-3),pp.11-17,1998.

[4] Y.Freund & R.E.Schapire : A decision-theoretic generalization of on-line learning and an application to boosting,Journal of Computer and System Sciences,55(1):119-139,1997.