

サンプリングによる ILP の効率的な学習

岩丸 悠一[†] 松井 藤五郎[‡] 大和田 勇人[‡]

東京理科大学大学院 理工学研究科 経営工学専攻[†] 同 理工学部 経営工学科[‡]

1. はじめに

機械学習の手法の一つである ILP のアルゴリズムは例(=事例, サンプル)数が高々数万程度と仮定して設計されている。

しかし, 現在はゲノムデータのように例数が遥かに大きなデータに対して機械学習を行なうことが求められており, この問題に対して学習するデータを削減するスケールダウン戦略が既存の学習アルゴリズムのまま用いることが出来ることから有効であると考えられている。

スケールダウン戦略には例数を減少させるデータサンプリング, 属性量を減少させる属性選択, データを要約し, 減少させるデータスカッシングなどが知られている。これらの手法を用いることで既存のアルゴリズムで学習が出来るが, 適用の仕方によってはデータの一般性が失われ, 導出される仮説の精度が低下してしまうという問題がある。特に, 属性を減少させることはデータに対する専門知識を要する場合があることや, ILP では仮説空間を事例の持つ情報から生成することからコストが大きいと考えられる。

そこで, 本研究では仮説空間に影響を与えずに既存のアルゴリズムで対応できるデータサンプリングを用いて ILP で大規模データを学習する手法を提案する。サンプリングを用いることによって1回の学習時間を減らしつつ, 精度低下を招かないよう, 繰り返し学習を行なう。さらに, 前回の学習結果に基づいた重み付きサンプリングと重み付き評価値を導入することで効率的な学習を行なう。

2. ILP

ILP とは事例とそれに関する情報(背景知識)から事例の一般化である仮説を導出する手法である。本節では, 代表的 ILP システムである GKS[1]に基づきその概要とサンプリングを利用する際の問題点を述べる。

Effective Learning for ILP using sampling

[†]Yuichi Iwamaru · Department of Industrial Administration, Graduate school of Science and Technology, Tokyo University of Science

[‡]Tohgoroh Matsui, Hayato Ohwada · Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

sample size(%)	accuracy(%)	runtime(s)
30	77	60
100	82	480

表1. サンプルサイズによる学習結果

2.1 仮説空間

GKS の仮説空間は逆伴意法によって形成される。訓練例集合から一つの事例を選択し, その事例を説明する最も特殊な仮説である最弱仮説を生成する。そして, 最弱仮説をボトムとして仮説空間を生成する。よって, 仮説空間は選択した事例の持つ背景知識にのみ依存し, 事例の数には関係しない。

2.2 仮説探索

GKS の仮説探索は MDL に基づいた最良優先探索を用いている。以下に P を仮説が説明した正事例数, N を説明した負事例数, Dep を仮説の変数深度, Len を仮説のリテラル数とした場合の MDL の求め方を示す。

$$MDL = P - N - Dep - Len$$

評価値は訓練例の事例数に依存する。そのため, サンプリングによって事例を減少させる場合, 仮説の評価が正しく行なわれないと考えられる。

2.3 サンプリングの利用

事例をサンプリングした場合の学習結果への影響を ILP のベンチマーク mutagenesis を学習した結果(表1)を基に示す。なお, 事例はランダムサンプリングによって選択している。

結果からサンプリングを行なうことで学習時間は大幅に削減されるが, 学習精度はそれほど低下しないと言える。よって, 繰り返し学習によって学習精度の低下を防ぐ手法を以下に述べていく。

3. 提案手法

本手法では, サンプリングを用いることによる精度低下を防ぐために繰り返し学習を行なう。その際に前回の学習結果に基づいた重み付きサンプリングと重み付き評価値を導入することで効率的に学習を行なう。

3.1 概要

事例集合 E が与えられたときの学習 I 回目の概要を以下に示す。なお, 1 回目の各事例の重みは $1/m$ (m =事例数)としている。

1. E から重み付きサンプリングにより訓練例集合 E_i を生成する.
2. E_i を訓練例集合として学習し, 学習器 L_i を生成する.
3. L_i で E を予測し, その精度に応じて L_i の評価を行なう.
4. 3の結果に応じて事例の重みを更新する
5. 1~4を n 回繰り返し, n 個の学習器を生成する.
6. 各学習器に 3 で得た評価に基づいた重み付き多数決により学習器を生成する.

以下に各ステップについて述べる.

Step 1. 重み付きサンプリング

事例の重みは学習の困難さを表す. 難しい事例を集中的に学習することでランダムに事例を学習するよりも効率的に学習を行なう. 特徴的な事例が多いため過学習の恐れがあるが, step 3 で事例全体での評価を行なうことでそれを防ぐ. 各事例の重みに関してはあとで示す.

Step 2. 重み付き評価値

前述したように仮説は難しい事例を説明している必要がある. よって, 仮説探索の評価値を重みが大きいものを説明するように変更する.

$$WMDL = P_w - N_w - Dep - Len$$

P_w, N_w は説明した正, 負事例の重みの総和である.

Step 3.4. 学習器の評価, 重みの更新

学習器の評価は事例全体に対する学習精度によって行なう. そして, その結果に応じて正しく説明した事例の重みは減少させ, 間違えて説明した事例の重みを増加させる.

Step 5.6 学習器の組み合わせ.

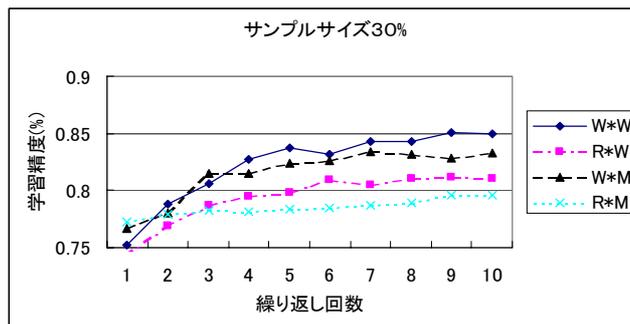
生成された複数の学習器はそれぞれの学習精度に応じた重みによって組み合わせ, 一つの学習器として生成する.

4. 実験

重み付きサンプリングと重み付き評価値 $WMDL$ の有効性を検証する. 比較としてランダムサンプリング, MDL を組み合わせ 4 通りの実験を行なう. データセットには *mutagenesis*, サンプルサイズは 30%, 繰り返し回数は 10 回とした. 結果は 10-fold cross validation を 10 回行った平均を示す(図 1, 表 2). なお, 図は {サンプリング方法*評価値} で表してあり, $W*W$ は重み付きサンプリングと評価値 $WMDL$ を用いた結果であることを示している.

5. 考察

学習精度は重み付きサンプリング, 重み付き評価値を用いた場合高い精度を得た. そして, その効果はサンプリングの方が強い結果を得た.



	重み付きサンプリング		ランダムサンプリング	
	WMDL	MDL	WMDL	MDL
学習精度	85%	83%	81%	80%
学習時間(s)	2095	1608	1391	1231

図 1, 表 2 実験結果

これにより, 学習に用いる事例が特徴的である方が学習に効率的に学習を行なえると言える. 学習時間は重み付きサンプリング, 重み付き評価値を用いた場合の方が多くかかった. これは難しい事例を集めているため問題が難しくなり一般化に時間がかかるためと考えられる.

単一で学習した場合(精度 82%, 時間 480s)と比較すると重み付きサンプリングを用いた場合は $WMDL$ の場合は 4 回目(82.7%, 624s)で, MDL の場合は 6 回目(82.6%, 641s)で精度を上回り, 10 回目は 85%, 83%と高い精度を得た. 上回った際の学習時間に関しては, 単一の場合と比べ $WMDL$ の場合 30%, MDL の場合 34%多かった. これはサンプルサイズ*繰り返し回数が $WMDL$ の場合 120%, MDL の場合 180%となるため, 本来の学習時間より上回ったと考えられる.

6. 結論

本論文では, 重み付きサンプリングと重み付き評価値を用いる繰り返し学習を行なうことで, 既存のアルゴリズムのまま大規模データを学習する手法を提案し, その効果を実験により示した. 重み付きサンプリングを用いることで単一の場合と同等の精度を得ることが出来, 特に重み付きサンプリングを用いる場合は繰り返し回数を増やすことで高い精度を得るという結果を得た. 今後の課題は同等の精度を得るまでの学習時間を改善することである.

参考文献

- [1] Ohwada, H., Nishiyama, H., and Mizoguchi, F.: Concurrent Execution of Optimal Hypothesis Search for Inverse Entailment, in *Proc. of ILP2000*, pp165-173, 2000.