1F2-02

重みつきサンプリングを用いたBoostingのILPへの適用

Application of Boosting using sampling with weight to ILP

岩丸 悠一[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Yuichi Iwamaru Tohgoroh Matsui Hayato Ohwada

[*1]東京理科大学大学院理工学研究科経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

[*2]東京理科大学理工学部経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

In the field of machine learning, the computational complexity when a large-scale problem is taken up becomes a problem. The application of Boosting is raised as a solution of this problem. In this paper, we proposed method of applying Boosting to ILP. The forcus is applied to the sample size of the learning. We used sampling with weight to reduce the examples size.

1. はじめに

機械学習の分野では, 大規模な学習モデルを構成する際の計算量の増大と汎化能力の低下が問題となっている. この問題を回避する手法の１つとして, 学習器を複数生成し, 単一の最終仮説として出力する Boosting アルゴリズムが注目を集めている. Boosting は学習アルゴリズムを複数回実行し, 各ラウンドで得られた仮説を統合して分類精度を改善する学習手法である. Boosting の代表的なアルゴリズムである AdaBoost は正答率が高いことが証明されている優れたアルゴリズムである.

しかし, Boosting の適用は命題学習においては注目されているが, ILP に対するアプローチはほとんど目をむけられていない. このことには２つの問題点が考えられる. 一つめの理由は ILP は表現力に優れた学習アルゴリズムであり, 複数の仮説を生成し, それらを組み合わせる Boosting では表現力を失ってしまうという点があげられる. そして, 二つ目の問題点としては Boosting は繰り返し学習することで単一の学習器を生成するため一回の学習時間が長い ILP に適用した場合, 実行時間が膨大になってしまうという点である. そのため, ILP への Boosting への適用は ILP の学習アルゴリズムの改善に目を向けられている [Hoche and Wrobel 01].

本研究では, ILP へ Boosting を適用する際に, 学習アルゴリズムではなく, 適用する際のサンプリングに注目をする. 本論文では, サンプルサイズの変化による学習精度と学習時間への影響を調査する. なぜなら, 後で示すように ILP の実行時間はサンプサイズの違いに大きく影響を受けるが, 学習精度に関してはサンプルサイズの違いに対してそれほど影響を受けない. そこで, 一回の学習にかかる時間を減少させるために学習に掛けるサンプルサイズを縮小する. サンプルサイズを小さくすることで各ラウンドにおける学習精度は低下するが Boosting の効果によって回復することができると考えられる. また, サンプリングの手法には重み付きサンプリングを用いる.

2. ILP

ILP(Inductive Logic Programming) は, 多くの事例とその背景知識から事例の一般化である仮説を導出する帰納推論を一階述語論理上でおこなうものである. 述語論理を枠組みとしているため, 豊かな表現力を持ち, より現実的な問題を扱えることで注目を集めている. しかし, ILP は探索の対象となる仮説空間が通常無限になるため学習速度が遅いことが知られている. 代表的な ILP システム GKS[Ohwada 00] は, 探索戦略に最良優先探索を用いている. 仮説空間は逆伴意法 [Muggleton 95] により有限内に制限されているが, 大規模なデータを扱う場合の仮説数は依然指数オーダとなるため, 学習時間の改善が必要とされている.

sample size(%)	accuracy(%)	runtime(s)
10	70	7.3
30	77	60.1
50	79	150
70	81	230
100	82	480

表 1: 学習精度と実行時間

3. Boosting

Boosting はアンサンブル学習の手法であり, 弱学習器によって構成された分類器を組み合わせることによって学習システムの予測精度を単一の強学習器のように向上させる手法である. AdaBoost[Freund and Schapire 97] とは Boosting の代表的なアルゴリズムであり, 与えられた学習アルゴリズムを用いて, 一回のラウンドで一個の学習器を生成する. 生成された仮説は訓練事例によってテストされ, 誤分類された訓練事例には高い重みが与えられる. 次回のラウンドでは, 重みを確率分布として訓練事例のサンプリングを行い, 得られた部分集合を新たな訓練事例として用いる. この操作によって, 各ラウンドで, それぞれ性質の異なる仮説が得られる. 最終的にこれらの仮説を１つの仮説に統合して, 高い分類精度を実現している.

4. BoostingのILPへの適用

Boosting を ILP へ適用する際の問題に学習時間が長いことがあげられる. 本論文では, ILP の学習時間が事例数に依存していることに注目し 1(データセットは mutagenesis を用いて一様なランダムサンプリングによって事例を選択), この問題を学習アルゴリズムの改善ではなく学習に掛ける事例数を減少させることで解決していく.

本研究では AdaBoost をベースとして次のようにして ILP システム GKS に Boosting を適用した. t 番目のラウンドにおける学習には, 入力として m 個の訓練事例 $(x_1, y_1), ..., (x_m, y_m)$ から各事例の重みを確率分布としたサンプリングを行い訓練例集合 D_t を生成する. x_i は事例集合 E の要素であり, y_i は x_i が正事例なら $+1$, 負事例なら -1 を取る. 各事例の重み $W_t(i)$ は, 訓練事例に対して, 重要度を反映して設定される. 重みの集合 W_t は正規化定数 Z_t によって正規化され $W_t(i) \in [0, 1]$ となる. また, 繰り返し回数 T は前もって決定する.

t 番目のラウンドにおいて学習された仮説を h_t とし, $h_t(x_i)$ は仮説 h_t が事例 x_i を正事例と予測したら $h_t(x_i) = +1$, 負事例と予測したら $h_t(x_i) = -1$ とする. 初期値として重みは均等に設定される. よって, すべての事例の重みは $1/m$ となる. また, t ラウンドにおける重みの更新規則は以下の式で与えられる.

$$D_{t+1}(i) = \frac{D_t(i)\exp(\alpha_t y_i h_t(x_i))}{Z_t}$$

各ラウンドで, 正しく予測された事例の重みは軽減し, 誤分類された事例の重みは増加させる. このため, 学習器は重みの大きい事例を集中的に学習する. また, h_t を得たあと仮説の重要度 α_t を以下の式に従って設定する.

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right).$$

ただし,

$$r_t = \sum_i D_t(i) y_i h_t(x_i)$$

である. すべてのラウンドが終了したあと, 得られたすべての仮説を結合して最終仮説 H を得る. H は α_t を h_t の重みとして用いた重みつき多数決によって得られる.

また, GKS で学習する際の評価関数 $MDL = P - N - Dep - Len$ を変更する.

$$MMDL = w_+ - w_- - Depth - Len$$

P は被覆した正事例数, N は被覆した負事例数, Dep は仮説に含まれる変数の深度, Len は仮説に含まれるリテラルの数を表す. w_+ は仮説に被覆された正事例の重みの総和. w_- は仮説に被覆された負事例の重みの総和である.

5. 実験と考察

5.1 実験

実際に本手法を ILP のベンチマークデータ mutagenesis に適用した結果を示す. サンプリングの手法には重み付きと一様なランダムを用いて, 繰り返し回数は 10 とする. サンプルサイズは 10%, 30%, 50%, 70%で行い 10-fold cross-validation にて評価を行った.

5.2 考察

サンプルサイズごとの学習時間を比較するとサンプルサイズを減少させることで飛躍的に学習時間を減少することができている. 学習精度に関しても低下は見られるが学習時間の減少と比較すると影響は小さいものであり, 全てのデータを用いて一度学習をした際の accuracy82% と比較してもサンプルサイズが 30% 以上 (ランダムサンプリングの場合は 50%) の場合は学習精度は向上している.

重み付きサンプリングとランダムサンプリングを比較すると学習時間はランダムサンプリングの方が短い. これは, 重みつきサンプリングで学習の難しい事例を集めるとそれらを一般化する仮説を探すための計算量は通常より多いものとなることからと考えられる. 小さいサンプルサイズ (30%, 50%) の場合は重みつきサンプリングの方が高い精度を得て, 大きいサンプルサイズ (50%, 70%) の場合はランダムサンプリングの方が優れた結果を得ている. これは少ないサンプルの場合は重みつきサンプリングによって学習すべき事例を効率よく選択できるが, 大きくなるとサンプリング手法に差が表れないためと考える. また, 10% の場合に全ての事例を用いた精度 (82%) より低くなっているのはサンプルサイズを小さくし過ぎたため, 仮説の一般性が下がったためと考えられる. また, サンプルサイズ 50% の時の重みつきサンプリングの時の学習精度, 学習時間ともに性能の低下が著しいがこれに関しては検証がさらなる実験と検証が必要である.

	weight_sampling		rundom_sampling	
size(%)	accuracy(%)	time(s)	accuracy(%)	time(s)
10	78	360	73	197
30	86	1486	78	1027
50	83	4149	91	3128
70	88	5964	88	5840

表 2: 実験結果

6. 結論

本論文では, ILP に Boosting を適用する際のサンプルサイズを変化させることでの学習時間と学習精度への影響を検証した. サンプルサイズを減少させた場合の学習時間の減少は大きく, また, それによって生じる学習精度の低下も Boosting を用いることで解決された. 重み付きサンプリングの利用はサンプルサイズが小さいほど有効であるが, サンプルサイズが小さ過ぎると仮説の一般性が損なわれるという問題もあるため, 今後も検証が必要である.

参考文献

[Muggleton 95] Muggleton, S.:Inverse Entailment and Progol, *New Generation Computing*, Vol.13, pp.245-286(1995)

[Freund and Schapire 97] Yoav Freund and Robert 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,August.

[Ohwada 00] Ohwada, H., Nishiyama, H., and Mizoguhi, F.: Concurrent Execution of Optimal Hypothesis Search for Inverse Entailment, in *Proc. of 10th International Conference on Inductive Logic Programming, Lecture Notes in Artificial Intelligence, SpringerVerlag No. 1866*, pp165-173,2000.

[Hoche and Wrobel 01] Susanne Hoche and Stefan Wrobel.:Relational Learning Using Constrained Confidence Rated Boosting. Proceedings of the 11th International Conference on Inductive Logic Programming volume 2157 of Lecture Notes in Artificial Intelligence, pages 51-64 Springer-Verlag, September 2001.