3M-1

ILP を用いた誤分類の判別による分類器の精度向上

横山 正樹 † 松井 藤五郎 ‡ 大和田 勇人 ‡

東京理科大学 理工学研究科 経営工学専攻 † 東京理科大学 理工学部 経営工学科 ‡

1 はじめに

事物をその性質に基づいてカテゴリに分類する分類問題では, 100%の精度を持つ分類器を作成することは難しい. すなわち, どのような分類器も誤分類を起こす. これは, 新規に観測された事例を分類するときによく見られる.

本研究では, この誤分類に着目し, 誤分類という概念を学習する問題としてとらえることによって誤分類を減らす手法を提案し, 分類精度を向上させる条件を示す. 本論文では, 概念学習の手法として ILP を使用する.

2 誤分類

本論文では二クラス分類問題について考える. 二クラス分類器を学習する問題は, 一方のカテゴリに着目すれば, 「そのカテゴリに含まれるという概念」を学習する問題とみなすことができる.

図 1のように, 誤分類は (a) 真のカテゴリが c ではないのに分類器が c に分類してしまったものと (b) 真のカテゴリが c であるのに分類器が c に分類しなかったものに分けられる. 前者 (a) を誤被覆事例 (miscovered examples), 後者 (b) を未被覆事例 (uncovered examples) と呼ぶ.

そこで, 本手法では, 与えられた分類器が生じる誤被覆と未被覆のそれぞれについて概念学習をおこない, 誤被覆事例あるいは未被覆事例を判別するルールを学習する. それぞれの概念を m, u と書く.

与えられた分類器がそのカテゴリに含まれると予測した事例を表す概念を $\hat{c}$ と書くとき, 本手法は x の分類を次のように予測する.

1. $\hat{c}(x)$ が真 かつ $\hat{m}(x)$ が真ならば, 偽に分類する.
2. $\hat{c}(x)$ が真 かつ $\hat{m}(x)$ が偽ならば, 真に分類する.
3. $\hat{c}(x)$ が偽 かつ $\hat{u}(x)$ が真ならば, 真に分類する.
4. $\hat{c}(x)$ が偽 かつ $\hat{u}(x)$ が偽ならば, 偽に分類する.

3 Miscovered Example の判別

まず, 誤被覆の概念を学習し, 与えられた分類器の誤被覆を修正する.

与えられた分類器の分類ルールを $\hat{r}$ とおく. そして, $\hat{r}$ によって「真」と分類された事例集合を抽出する. その中で, 図 2のように FP_r の事例 (miscovered examples) を正事例, TP_r の事例を負事例とおき, ILP を用いて miscovered example 判別ルール $\hat{m}$ を学習する.

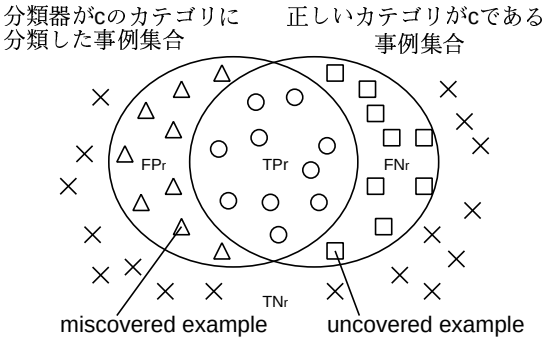

図 1: Miscovered example と uncovered example

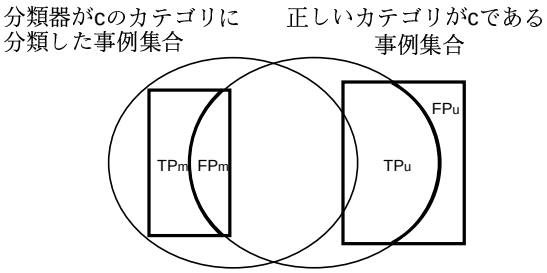

図 2: Miscovered example と uncovered example の判別

$\hat{m}$ に被覆された事例は誤分類と判別される. したがって, $\hat{m}$ に被覆された事例を「偽」に再分類する. ここで, $\hat{m}$ により正しくカテゴリを再分類できた集合を TP_m とし, 誤ったカテゴリに再分類してしまった集合を FP_m とする. そして, miscovered example の判別と修正を行った集合は

$$\mathrm{TP}_{rm} = \mathrm{TP}_r \setminus \mathrm{FP}_m \qquad \mathrm{FP}_{rm} = \mathrm{FP}_r \setminus \mathrm{TP}_m$$
$$\mathrm{FN}_{rm} = \mathrm{FN}_r \cup \mathrm{FP}_m \qquad \mathrm{TN}_{rm} = \mathrm{TN}_r \cup \mathrm{TP}_m$$

となる.

4 Uncovered Example の判別

次に未被覆の概念を学習しする.

r によって「偽」と分類された事例集合を抽出する. その中で, 図 2のように FN_r の事例 (uncovered examples) を正事例とおき, TN_r の事例を負事例とおく. また, uncovered example だけを被覆するルールを学習するために TP_r の事例も負事例に追加する. すなわち FN_r の事例を正事例, TN_r, TP_r の事例を負事例とおき, ILP を用いて uncovered example の判別ルール $\hat{u}$ を学習する.

$\hat{u}$ に被覆された事例は誤分類と判別される. したがって, $\hat{u}$ に被覆された事例を「真」に再分類する. ここで, $\hat{u}$ により正しく分類できた集合を TP_u とし, 誤ったカテゴリに再分類してし

Improving accuracy of a classifier by discriminating errors using ILP

Masaki YOKOYAMA†, Tohgoroh MATSUI‡ , Hayato OHWADA‡

†Department of Industrial Administration, Graduate school of Science and Technology, Tokyo University of Science ‡Department of Industrial Administration, Faxulty of Science and Technology, Tokyo University of Science

278-8510, Noda, Japan

まった集合を FP_u とする．そして, uncovered example の判別と修正を行った集合は

$$TP_{ru} = TP_r \cup TP_u \qquad FP_{ru} = FP_r \cup FP_u$$
$$FN_{ru} = FN_r \setminus TP_u \qquad TN_{ru} = TN_r \setminus FP_u$$

となる．

Miscovered example と uncovered example の両方を判別し修正した提案手法の集合は

$$TP_{rmu} = (TP_r \setminus FP_m) \cup TP_u \quad FP_{rmu} = (FP_r \setminus TP_m) \cup FP_u$$
$$FN_{rmu} = (FN_r \cup FP_m) \setminus TP_u \quad TN_{rmu} = (TN_r \cup TP_m) \setminus FP_u$$

となる．

5 提案手法における定理

本手法においては, 次のような定理が成り立つ．スペースの都合により証明は省略する．ここで, Recall, Precision, Accuracy をそれぞれ r, p, a とおく．

定理 1 : $r_{rm} \leq r_r$.
定理 2 : $p_m \geq 1 - p_r$ならば $p_{rm} \geq p_r$. また, その逆も成り立つ.
定理 3 : $p_m \geq 1/2$ ならば $a_{rm} \geq a_r$. また, その逆も成り立つ.
定理 4 : $r_{ru} \geq r_r$.
定理 5 : $p_u \geq p_r$ならば $p_{ru} \geq p_r$. また, その逆も成り立つ.
定理 6 : $p_u \geq 1/2$ ならば $a_{ru} \geq a_r$. また, その逆も成り立つ.
定理 7 : $r_{rmu} \geq r_{rm}$.
定理 8 : $p_u \geq p_{rm}$ならば $p_{rmu} \geq p_{rm}$. また, その逆も成り立つ.
定理 9 : $p_u \geq 1/2$ ならば $a_{rmu} \geq a_{rm}$. また, その逆も成り立つ.
定理 11 : $p_m \geq 1 - p_r$かつ $p_u \geq p_{rm}$ならば $p_{rmu} \geq p_r$.
また, その逆も成り立つ.
定理 12 : $p_m \geq 1/2$ かつ $p_u \geq 1/2$ ならば $a_{rmu} \geq a_r$.
また, その逆も成り立つ.

6 実験

ここで, 提案手法の有効性を示すために実験を行った. データはペンシルバニア大学の Penn TreeBank Project によって作成された Wall Street Journal の記事を使用した. このコーパスに含まれる単語は 45 種類の品詞・記号に分類されている. ここでは英単語への品詞タグ付け問題に本手法に適用した.

本実験では, 分類器として Brill's Tagger [1] を与え, ILP システムの GKS [2] を用いて誤分類判別ルールを学習した. ここで, 背景知識はその単語と前後 3 つの単語とタグを使用した. また, 事例は miscovered example について学習するときは $FP_r : TP_r$ を 1:1, uncovered example について学習するときは $FN_r : TN_r : TP_r$ を 1:1:1 の割合でランダムサンプリングを行い, 10-fold cross validation で評価した. ここで, uncovered example では TP_r を削除した FN_r と TN_r をテストセットとした.

評価方法は Brill's Tagger のみの Recall (r_r), Precision (p_r), Accuracy (a_r) と提案手法の Recall (r_{rmu}), Precision (p_{rmu}), Accuracy (a_{rmu}) を比較した.

表 1: Brill's Tagger のみの Racall (r_r), Precision (p_r), Accuracy (a_r) と提案手法の Recall (r_{rmu}), Precision (p_{rmu}), Accuracy (a_{rmu})

Brill's Tagger			提案手法		
r_r	p_r	a_r	r_{rmu}	p_{rmu}	a_{rmu}
0.9501	0.9502	0.9978	0.9502	0.9531	0.9986

表 1は Brill's Tagger と提案手法を比較したものである. Brill's Tagger のみよりも提案手法のほうが精度が高かった. また, タグごとに評価した場合, 21 個のタグで Recall, Accuracy を上げることができた. Precision は 17 個のタグで上げることができた. しかしながら, 4 個のタグで提案手法のほうが下がってしまった.

また, 前置詞について獲得したルールを一部を以下に掲載する.

```
miscovered(A,B):-post1word(A,B,'.').    (1)
uncovered(A,B):-word(A,B,'up').         (2)
```

(1) は「次の単語が "." であるときその単語は miscovered example である」. (2) は「その単語が "up" であるときその単語は uncovered example である」ことを意味している.

7 考察

4 つのタグで提案手法のほうが Precision が低くなってしまった．ここで定理 11 より, uncovered example の判別ルールの Precision (p_u) が miscovered example を判別し修正した Precision (p_{rm}) を超えていなければ精度は上がらない. そして, p_{rm} がすでにかなり高いので, p_u はそれ以上に高くならなかった. したがって, 提案手法のほうが Precision が下がってしまったのは, 定理 11 が示すとおりの結果であると言える.

また, 獲得した uncovered example の判別ルールは Brill's Tagger に組み込むことにより Brill's Tagger そのものの精度を向上させることができるだろう.

8 まとめ

本論文では, ILP を用いて誤分類に着目し, 誤被覆・未被覆という二つの概念を学習する問題としてとらえることによって誤分類を減らす手法を提案した. また, 分類精度を向上させる条件を示した.

実験では全体の精度を向上させることができた. タグごとの評価では, 提案手法は 21 個のタグで Recall, Accuracy を向上させた. Precision は 4 個下がってしまったが, 17 個上げることができた. したがって, 本手法は有効だと言える.

参考文献

[1] Eric Brill. Some advances in transformation-based part of speech tagging. In *Proc. of AAAI-94*, 1:722–727, 1994.

[2] Fumio Mizoguchi and Hayato Ohwada. Constrained relative least general generalization for inducing constraint logic programs. *New Generation Computing*, 13:335–368, 1995.