2F3-03

ブール関数の学習におけるブーリアンカーネルを用いた特徴選択について

Feature selection using Boolean kernels for the learning of Boolean funtions

佐土原 健[*1]
Ken SADOHARA

[*1]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

This paper considers a variable selection algorithm that learns classifiers of bit-vectors and identifies useless variables for the classification by analyzing the classifiers. By using Boolean kernels, it learns the linear thresholding hypothesis over conjunctions of given variables and, for each variable, it computes the square sum of weights of conjunctions containing the variable. Then the variable yielding the smallest square sum is identified as the most useless variable. It is shown that the algorithm outperforms several existing algorithms in experiments on artificial datasets and a dataset for text categorization.

1. はじめに

近年, データマイニングが対象とするデータは, 例えば, マイクロアレイ分析やテキスト分類等で用いられるデータのように, 非常に多くの変数で記述されている場合が多い. そうした状況の下で, データの分析に寄与する特徴 (変数) を選択する方法に関する研究が, 再び注目を集めている [3]. 特に, 分類学習における変数選択は, データを記述する多くの変数の中から, 分類に寄与する変数を選択する問題である. 分類にとって十分な少数の変数を選択することは, 分類精度の向上が期待できるだけでなく, 学習や分類に必要な計算資源を節約したり, データをより良く理解するためにも有用である.

本研究では, ブール関数の学習における変数選択について考察する. その理由は, 離散データに対する分類学習は, 本質的にブール関数の帰納学習の問題に帰着されるからである. さらに, 数値データに対しても, 可読性や計算資源の節約のために, 変数が離散化される場合も多いことを考えれば, ブール関数の学習における変数選択を研究することの意義は大きい.

変数選択においては, 変数間の依存関係を考慮に入れる必要がある. 例えば, ブール式 $y = x_1\overline{x_2} \vee x_2x_3$ において, x_2 は y にとって必要不可欠な変数であるにも関わらず, x_2 の値を知ることは, 全く情報利得をもたらさない. したがって, クラス変数との相互情報量により x_2 を評価すると, x_2 は分類に不必要な変数であると判断されてしまう. このような変数の依存関係を考慮するために, 本研究では, Support Vector Machine (SVM) [1] を用いた, 変数選択アルゴリズムについて考察する.

このアルゴリムは, 文献 [3] で提案された, Recusive Feature Elimination (RFE) に基づいて次のように動作する. まず, SVM を用いて, 論理積が張る空間上で, 論理積の線形和としてブール関数を学習する. そのような空間は一般に非常に高次元であるが, ブーリアンカーネルを用いることで, 効率良くブール関数を学習できることが知られている. こうして, 得られた論理積の線形和に対して, 特定の変数を含む全ての論理積の重みの二乗和を計算し, そのような二乗和が最も小さい変数を, 最も分類に寄与しない変数と判断する. 一般に, 特定の変数を含む全ての論理積の数は非常に多いが, ブーリアンカーネルを用いることで, このような論理積の重みの二乗和を効率良く計算することが可能である. 本研究では, このようなアルゴリズムが, 人工的に生成したデータセットと, テキスト分類のベンチマークデータセットを用いた計算機実験において, 既存の変数選択アルゴリズムよりも優れていることを示す.

連絡先: 佐土原健, sadohara@computer.org, 〒 305-8568 つくば市梅園 1-1-1 つくば中央第二, Tel: 029(861)5922

2. SVM とブーリアンカーネル

SVM は, 与えられた訓練データ $\boldsymbol{x}_i \in X, y_i \in \{+1, -1\}$ $(i = 1, \ldots, n)$ に対して, 特徴空間上のデータ $\phi(\boldsymbol{x}_i)$ を正しく分離できる最大のマージン持つ超平面 $f(\boldsymbol{x}) = \langle \boldsymbol{w} \cdot \phi(\boldsymbol{x}) \rangle + b = 0$ を学習する. この最大マージン超平面は, 次の最適化問題

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{subject to} \quad \sum_{i=1}^{n} y_i \alpha_i = 0, \quad \alpha_i \geq 0 \quad (1 \leq i \leq n).$$

を解くことで, $f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b$ のように得られる. ここで, K は, 特徴空間の内積 $\langle \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) \rangle$ を計算する関数であり, カーネル関数と呼ばれる. カーネル関数を用いることで, 一般に計算が困難な ϕ を陽に計算することなく, 特徴空間上の最大マージン超平面の学習が可能になる.

本研究では, ブール関数の学習のために, 論理積の張る特徴空間を考えるが, このような空間に対する次のようなカーネル関数が知られている [5]. 長さが高々 k である論理積が張る特徴空間に対しては, $K^k(\boldsymbol{u}, \boldsymbol{v}) \stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathrm{a}(\boldsymbol{u},\boldsymbol{v})\mathrm{C}_i$. 否定を含まない長さが高々 k である論理積が張る特徴空間に対しては, $K^k_m(\boldsymbol{u}, \boldsymbol{v}) \stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathrm{p}(\boldsymbol{u},\boldsymbol{v})\mathrm{C}_i$. ここで, $\mathrm{a}(\boldsymbol{u}, \boldsymbol{v})$ は, ビット列 $\boldsymbol{u}$ と $\boldsymbol{v}$ において同じ値を持つビットの数を表わし, $\mathrm{p}(\boldsymbol{u}, \boldsymbol{v})$ は, $\boldsymbol{u}$ と $\boldsymbol{v}$ において共に値 1 を持つビットの数を表わす. これらのカーネル関数が, 特徴空間の次元に依存せずに, 効率良く計算可能であることに注意されたい.

3. Recursive Feature Elimination

RFE では, このように学習された超平面に対して, 各変数 x ごとの評価値 $D(x) = \sum\sum \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum\sum \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i(-x), \boldsymbol{x}_j(-x))$ を計算する. ここで, $\boldsymbol{u}(-x)$ は, ベクトル $\boldsymbol{u}$ から x に対応する成分を取り除いたベクトルを表わす. 最適解 $\alpha_1, \ldots, \alpha_n$ に対して, $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$ であるので, K として, ブーリアンカーネルを用いる場合, $c(x)$ を x を含む論理積の添字の集合とするとき, $D(x) =$

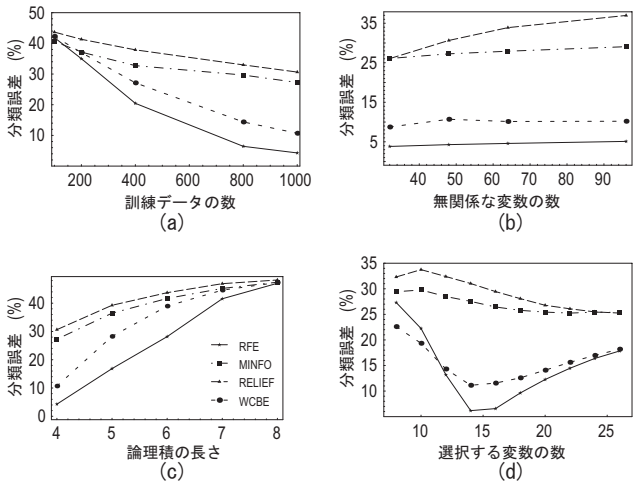

図 1: 人工データにおける性能比較

$\sum_{\ell \in c(x)} w_\ell^2$ となる. RFE は $x^* = \mathrm{argmin}_x D(x)$ を, データから削除した後で, 超平面を再学習し, このようなプロセスを繰り返すことで, 分類に寄与しない変数を次々に除去していく.

4. 実験:人工データ

この実験では, 人工的に合成されたブール関数の入出力例の集合から, 各変数選択アルゴリズムにより選択された変数のみを用いてブール関数を学習し, その分類誤差を測定することにより, 次の 4 つの変数選択アルゴリズムの性能を比較する: (1)RFE, (2)MINFO (相互情報量を用いた変数ランキング法), (3)RELIEF, (4)WCBE (C4.5 を用いた wrapper 法 [4]).

データの生成は, 3 つのパラメタ (1)DNF 式の真偽に無関係な変数の数 r, (2) 訓練データの数 n, (3) 論理積の長さ ℓ で定義される DNF 式の複雑さ, を制御して, 以下のように行われる. まず, $16 + r$ 個の変数の中で, ある固定された 16 個の変数のみを用いて DNF 式を生成する. DNF 式の各論理積は, ランダムに選ばれた ℓ 個の変数を $\frac{1}{2}$ の確率で負リテラルとすることで生成される. 論理積の数は, $2^{\ell-1}$ 個とする. そして, この DNF 式に対して, n 個の訓練データと, 2000 個のテストデータが, 一様分布の下で独立に生成される.

このように生成されたデータは, 各変数選択アルゴリズムに与えられ, m 個の変数が選択される. 次に, 選択された変数とデータから, 共通の学習アルゴリズム (K^ℓ を用いた SVM) により分類器を学習し, テストデータに対する分類誤差を測定する. このような測定を, 160 個の DNF 式に対して行い, その平均値を用いて変数選択アルゴリズムの性能を比較した. 図 1(a) は, $\ell = 4$, $r = 48$ のときに, n の変化に対する分類誤差の変化を表わしている. 図 1(b) は, $\ell = 4$, $n = 1000$ のときに, r の変化に対する分類誤差の変化を表わしている. 図 1(c) は, $r = 48$, $n = 1000$ のときに, ℓ の変化に対する分類誤差の変化を表わしている. 以上の実験では, m は 各 DNF 式に表われた変数の数としたが, 図 1(d) は, $r = 48$, $n = 1000$, $\ell = 4$ のときに, m を変化させた実験の結果である.

5. 実験:テキスト分類

実データに対する変数選択アルゴリズムの性能を比較するために, テキスト分類のデータセットである Reuter-21578 を用いた実験を行った. 実験には, 文献 [2] で用いられた, 前処理済

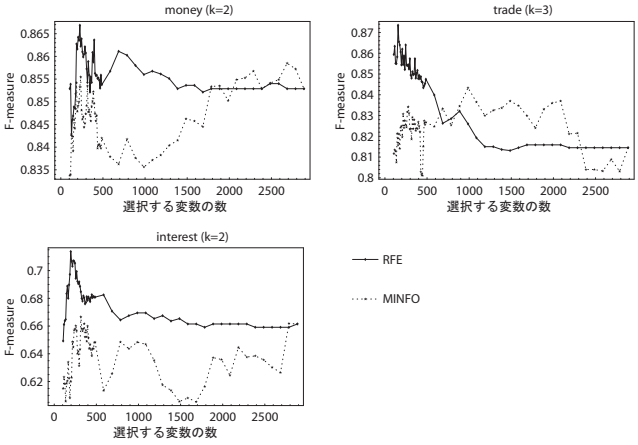

図 2: テキスト分類データにおける性能比較

みのデータセットのうち, "re0" と呼ばれるデータセットを用いた. このデータセットには, 1504 のニュース記事が含まれ, 各記事は, 分類カテゴリが付与された, 2886 次元の 2 値ベクトルで表現されている. 図 2 は, 最も正例の多い 3 つのカテゴリ "money", "trade", "interest" の F-measure 値の変化を示している. この実験で用いた RFE は, d 個の変数が残っているときに, 一度に $10^{\lfloor \log_{10} d - 1 \rfloor}$ 個の変数を除去する. また, 学習アルゴリズムとして K_m^k を用いた SVM を使用して, 8-分割交差検定により F-measure 値を求めた.

6. まとめ

これらの実験から, ブーリアンカーネルを用いた RFE が, 分類に寄与しない変数を除去することで, 高い分類精度をもたらし得ることが分る. 特に, テキスト分類の実験結果は, 変数間の相互作用を考慮に入れない手法に比べて, 少ない変数で高い分類精度を達成し得ることを示している.

謝辞 本研究は, 科研費 若手研究 (B)(No.14780315) の支援を一部受けている.

参考文献

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge Press, 2000.

[2] G. Forman. An extensive empirical study of feature selection metrics for text classification. *JMLR*, 3:1289–1305, 2003.

[3] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3:1157–1182, 2003.

[4] G.H.John, R.Kohavi and K.Pfleger: "Irrelevant features and the subset selection problem", Proc. of ICML, pp. 121–129, 1994.

[5] R.Khardon, D.Roth and R.Servedio. Efficiency versus convergence of Boolean kernels for on-line learning algorithms. NIPS 14:423–430, 2002.

[6] K. Sadohara. On a capacity control using Boolean kernels for the learning of Boolean functions. Proc. of ICDM, pp.410–417, 2002.