文字列集合の分類問題における最適パタン発見の効率的アルゴリズム

数理情報学専攻 48156226　守屋 航一
指導教員　定兼 邦彦 教授

1　序論

多くのデータ・変数を扱う機械学習において，メモリ容量は常に付きまとう問題であり，初期からメモリ容量に着目した工夫・改良が行われている．カーネル法など入力から非線形的特徴を抽出しようとする場合，特徴に対応した変数すべてがメモリに載らなくなり，変数の更新に工夫を要する．データそのものがメモリに載らない場合，ディスクのデータにアクセスする時間が問題となってくる．データがメモリに載らない場合のアプローチとして，Langford らは 2 つの方向性，並列化とオンライン学習を上げている [2].

このような状況における正則化項付き最小化問題に対するブロック化最小化スキームは，Yu らによって初めて提案された [3]．この手法では，データをいくつかのブロックに分解し，そのブロックに含まれるデータのみで変数を更新している．Matsushima は座標降下法を用いたスキーム Feature Cashed Loop (FCL) を提案し，並列可能を可能にしつつ，各変数の更新に用いるメモリ容量を抑えることに成功した [1]．一方で，このスキームにおいて，更新する変数はランダムに選択されているが，組合せの数が膨大であるため，目的関数の最適化に寄与しない変数を選んでしまうという問題が存在した．本研究は，この変数選択に関する高効率のアルゴリズムを提案するものである．

2　学習モデルと最適化手法

入力となるデータ列 $X = (x_i), (i = 1, \ldots, n)$ から出力となるデータ列 $\mathbf{y} \in \mathcal{R}^n$ を予測する問題を考える．計算のため，データ $x_i \in \mathcal{X}$ を特徴関数 $\phi : \mathcal{X} \to \mathbb{R}^p$ を用いて p 次元ベクトルに写像する．このとき，$\mathbf{w}^{\mathrm{T}}\phi(X) \approx \mathbf{y}$ となるパラメータベクトル $\mathbf{w} \in \mathbb{R}^p$ を求める問題を考える．FCL は以下の式 1 で表される関数の最小化として定式化される L1 ロジスティック回帰の学習を座標降下法を用いて実装している．

$$L(\mathbf{w}) = \|\mathbf{w}\|_1 + C\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{w}^{\mathrm{T}}\phi(x_i))) \quad (1)$$

座標降下法は更新するパラメータの 1 つの要素に着目し，その 1 要素のみを変数とする最適化問題と見て，その最適値で更新する，という手法であり，1 つのパラメータの更新に対してそのパラメータのみを用いて行うことができる．L1 ロジスティック回帰では近似を用いることで，以下の更新式が導出される．

$$w_j^{t+1} = \begin{cases} w_j^t - \frac{D(\mathbf{w},\mathbf{x}_i,y_i)+\frac{1}{2C}}{\sum_i \phi_j^2(\mathbf{x}_i)} & w_j^t > \frac{D(\mathbf{w},\mathbf{x}_i,y_i)+\frac{1}{2C}}{\sum_i \phi_j^2(\mathbf{x}_i)} \\ w_j^t - \frac{D(\mathbf{w},\mathbf{x}_i,y_i)-\frac{1}{2C}}{\sum_i \phi_j^2(\mathbf{x}_i)} & w_j^t < \frac{D(\mathbf{w},\mathbf{x}_i,y_i)-\frac{1}{2C}}{\sum_i \phi_j^2(\mathbf{x}_i)} \\ 0 & \text{o.w.} \end{cases}$$

$$D(\mathbf{w}, \mathbf{x}_i, y_i) = \sum_i (\mathbf{w}^{t\mathrm{T}}\phi(\mathbf{x}_i) - y_i)\phi_j(\mathbf{x}_i)$$

FCL では，writer と trainer の 2 つのスレッドを独立して動かす．writer は，ディスクのデータを読み取り，ランダムに選ばれた $\phi_j(\mathbf{x}_i)$ をキャッシュに書き出す．一方 trainer は，キャッシュからランダムに選んだ特徴のパラメータを上の更新式を用いて更新する．2 つのスレッドは独立して動かすことが可能であり，並列実行可能であることがこのスキームの強みの 1 つといえる．

3　実問題とその定式化

本研究で扱う実問題は，スプライス部位認識問題と呼ばれるバイオインフォマティクスにおける重要な問題である．機械学習における 2 クラス分類問題として解釈できる問題であり，入力 $\mathbf{x}_i \in \Sigma^d$ と出力 ± 1 の組から，出力が $+1$（または -1）に共通するパタンを見つける問題といえる．このとき，パタンにはワイルドカード（任意の文字と一致する記号）を許すものとする．この問題を正則化項付きリスク最小化問題として式 1 のように定式化し，前述のアルゴリズムおよびスキームを用いて学習を行う．

4　最適パタン発見問題とそのアルゴリズム

FCL スキームでは，更新するパタンをランダムに選択していたが，パタン数の多さからその効率性に問題があった．また FCL スキームでは，「$|C\nabla_j L(\mathbf{w}^t)|$ の値が小さいものは 0 に収束する」というヒューリスティクスを利用して，パタンの選別を行っていた．そこで，この値をスコアとして，全パタンの中からスコアが閾値

以上のパタンを列挙する問題を考える．パタンの探索方向として「ワイルドカードの少ないものから調べる」Upward Algorithm と「ワイルドカードの多いものから調べる」Downward Algorithm の 2 つを提案した．

Upward Algorithm はすでに計算したパタンのスコアの足し合わせにより次のパタンのスコアを計算することができるが，現れた全てのパタンを探索する必要がある．一方で，Downward Algorithm ではスコアの計算に毎回パタンの探索を行う必要があるものの，適宜条件に応じた探索の打ち切りを行うことができる．

5 実験

既存手法と提案手法の比較および提案 2 手法の比較を実験で行った．

既存手法と Upward Algorithm の比較では，時間ごとの目的関数値の減少度・学習性能を表す AUPRC の増加度いずれにおいても提案手法が優っていた．また，正例・負例比率の偏った元データではテスト精度の改善度合いが測れなかったため，正例・負例を同数にした人工データに対して同様の実験を行い，テスト精度においても，既存手法が 0 であり続けたのに対し，提案手法では 0.8 まで短時間の計算で得られることを示した．

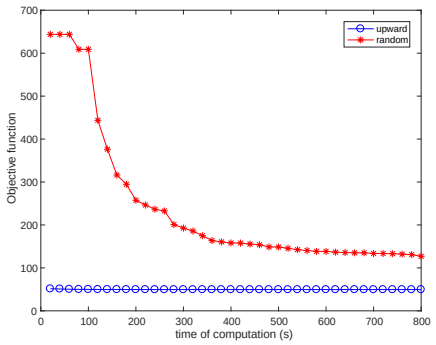

図 1. 既存手法と U.A. の比較．20s ごとの目的関数値

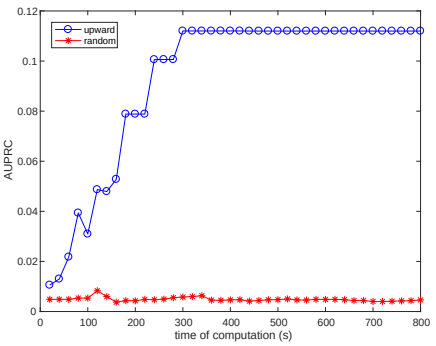

図 2. 既存手法と U.A. の比較．20s ごとの AUPRC

提案 2 手法をパタン長 b を変えて比較したところ，$b = 8$ の場合 Upward Algorithm が計算速度で優っていたが，$b = 12$ では Downward Algorithm が優る結果が得られた．パタン長が短い場合，Upward Algorithm が優位であり，長い場合に Downward Algorithm の枝刈りの優位性が現れることを示した．

パタン長を変えて Downward Algorithm を計算した結果を比較し，短パタン長でのスコア上位パタンが長パタン長でのスコア上位パタンに現れることを示した．また，長パタン長でのみ表現可能なパタンが上位に出現することもわかった．短時間の計算での比較では，パタン長が短い場合に計算速度が高いことから，最適化速度および学習速度で優位になることを示した．一方，長時間の実験から，パタン長が長い場合でも学習度合の改善が図れることもわかった．

ワイルドカード数に制限を加えて行った実験からは，制限によりパタン総数が少なくなることで計算速度向上は見込めるものの，表現力の低下から最適化・学習性能で良い結果を得ることはできなかった．

6 結論

本研究では，文字列集合分類問題に対する機械学習スキームに対し，最適なパタンを発見する効率的アルゴリズムを 2 つ提案した．このアルゴリズムを用いることで，従来手法よりも高速な最適化・学習を行うことに成功した．2 つのうち Upward Algorithm はパタン長が短い場合に特に有効であり，Downward Algorithm はパタン長が長くなっても十分な速度で計算が可能であることを実験により示した．

本研究で用いたデータは非常に正例の数が少なく，それに対応したアルゴリズムを構成したものの，正例に共通するパタンの発見は長時間の計算を経ても厳しいことがわかった．このため，さらなる高速化やパタンの拡張などが求められるだろう．

参考文献

[1] S. Matsushima. (2016). Asynchronous Feature Extraction for Large-Scale Linear Predictors. In Frasconi P., Landwehr N., Manco G., Vreeken J. (eds) *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2016. Lecture Notes in Computer Science, vol. 9851. Springer, Cham.

[2] J. Langford, L. Li and T. Zhang. (2009). Sparse online learning via truncated gradient. In *JMLR*, vol. 10, pp. 771–801.

[3] H.F. Yu, C.J. Hsieh, K.W. Chang, and C.J. Lin. (2010). Large linear classification when data cannot fit in memory. In *Proceedings of Knowledge Discovery and Data Mining*, pp. 833–842.