計算応用科学分野

ランキング上位の整合性に特化した効率的な情報検索アルゴリズム

学籍番号 23413565　氏名 松野 司
指導教員名 竹内一郎 准教授

1　まえがき

近年の情報技術の発達により Web 上に大量のデータが氾濫するようになり個人の求める情報を自力で見つけることが困難となっている．そのため, 検索エンジンのようなランキング技術は人々の生活にとって必要不可欠なツールとなっている．しかし, この分野で良く知られているサポートベクトルマシン (SVM) を用いたモデルでは, ユーザーのドキュメントへの参照頻度が上位数件に集中しているという特徴に適応されていなかった．そこで，本研究では, 各ドキュメントのランキング情報を学習に取り入れることによってランキング上位の整合性に特化した学習を行った．その際, 最適化問題が, 非凸最適化問題になるという欠点を抱えていたが, 部分的な凸計画問題を繰り返すことによって局所最適解に収束するアルゴリズムを提案した．同時に, 最適解における解の特徴を利用することによって学習を高速化する手法を提案した.

2　ランキング学習

ランキング学習の目的は，クエリと文章から抽出された特徴ベクトルに基づき，各文章のクエリに対するランキングを出力することである．ランキング学習の学習データは, クエリと文章から抽出された特徴ベクトル $\boldsymbol{x}_i \in \mathbb{R}^p$ と被験者により判定された関連度 $y_i \in \{0, 1, \ldots\}$ のペア集合 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ により構成される．ランキング学習では, 特徴ベクトル $\boldsymbol{x}$ の関連度の高いほど大きな値を出力する関数 $f : \mathbb{R}^p \to \mathbb{R}$ を作成することが目標となる．

関数 f を $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ とすると，ランキング学習は, 各インスタンスのペア $(i, j), y_i \neq y_j$ に関して,

$$\begin{aligned} \boldsymbol{x}_i \succ \boldsymbol{x}_j &\Leftrightarrow f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j) \\ &\Leftrightarrow \boldsymbol{w}^\top (\boldsymbol{x}_i - \boldsymbol{x}_j) \end{aligned}$$

となるパラメーター $\boldsymbol{w} \in \mathbb{R}^p$ を求めることになる．ただし，”$\succ$”は関連度の大小関係を表す．このような大小関係を適切に学習するため，各特徴ベクトルのペアを考え, 新たに $\{(\boldsymbol{x}_{i1} - \boldsymbol{x}_{i2}), z_i\}_{i=1}^n$ を作成する．ここで,

$$z_i = \begin{cases} +1, & y_{i1} > y_{i2} \\ -1, & y_{i1} > y_{i2} \end{cases}$$

であり, n は全ての可能なペア数を表す．この新たなデータを学習データとして 2 クラス分類問題を解くことによって上記の大小関係を満たす関数を学習することができる．その際に，同じペアで順序を入れ替えただけのペアは境界からの距離が等しいという特徴を利用し, $z = +1$ のペアのみ学習に利用し, そのペアの集合を D とする．

3　上位を重視した RankingSVM

RankingSVM では, ペアによって作成された 2 クラス分類用のデータを SVM に適用するものであるが, 各ドキュメントの関連度やドキュメントがランキングのどこに位置しているか, などの情報が欠落している欠点を抱えていた.

そこで, SVM のマージンの内側に入ってしまった際に与えられるペナルティの値を, ペア点を構成しているドキュメントとランキングトップのドキュメントとの出力の差だけ軽減することによって, 相対的に上位を重視した学習を行えるようにした．ランキングトップのインスタンスを k, g_{ij} を二つのドキュメントの出力の差 $f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)$ とすると, 提案法の目的関数は,

$$\min_{\boldsymbol{w}, k} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C \sum_{(i,j) \in D} [1 - g_{ij} - \lambda(g_{ki} + g_{kj})]_+, \quad (1)$$

となる．ただし, C と λ は非負のハイパーパラメーター, $[a]_+ = \max(a, 0)$ である．

4　局所最適解への収束

提案法では, ランキングトップ k との出力の差が定式化 (1) に含まれているが, どのドキュメントがトップになるのかを事前に知ることはできない．そのため，関数の学習とトップの同定を行う必要

があるのだが，両方を同時に行うと，問題が滑らかでない非凸最適化問題となり局所最適解を求めるのが困難となる．

しかし, k をいずれかのドキュメントに固定した状態であれば, 凸計画問題として定式化され, 最適解の導出が可能となる．そこで，k をいずれかのドキュメントで固定した条件付最適化問題:

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C \sum_{(i,j)\in D} [1 - g_{ij} - \lambda(g_{ki} + g_{kj})]_+,$$
$$\text{s.t.} \quad g_{k\ell} \geq 0 \quad \ell \in E$$

の条件付最適解を繰り返し算出することによって適切な k の同定と, 局所最適解の算出を行う．ただし，E はトップのドキュメントを除く同一クエリ内の全てのドキュメントの集合である．

条件付最適解が局所最適解の条件を満たせば, その解が局所最適解となる．条件を満たさない場合は, その条件付最適解を用いて, 次の条件付最適化問題を導出し, 再度, 条件付最適解を求める．この繰り返しによりアルゴリズムは確実に局所最適解に収束することが保証されている．

5　学習の高速化

提案法では, 双対問題を解くことによって条件付最適化問題を導出しているが, 双対問題における最適性 (KKT) 条件に着目すると, g_{ki} の値が十分に大きいドキュメントに関するラグランジェ未定乗数の最適解は, 非常に高い確率で 0 となる．

そこで，g_{ki} の値が閾値より大きい, つまり, ペア点を構成しているドキュメントの出力値が十分に小さく, ランキングの下位に位置している場合, そのドキュメントに関する最適化パラメーターの値を 0 に固定し, 最適化パラメーターから一時的に外すことができる．それにより, 最適化パラメーターの数を減らし，学習の高速化を実現した．

6　実験

ベンチマークデータセットとして, Yahoo! leaning to rank challenge (ICML2010) のデータセット ($\boldsymbol{x}$ は 700 次元, y は 5 段階 ($y \in \{0, 1, \ldots, 4\}$)) を用いた．

学習時間の比較 (単位:秒)	
従来法 (Past)	提案法 (Proposal)
2489	1365

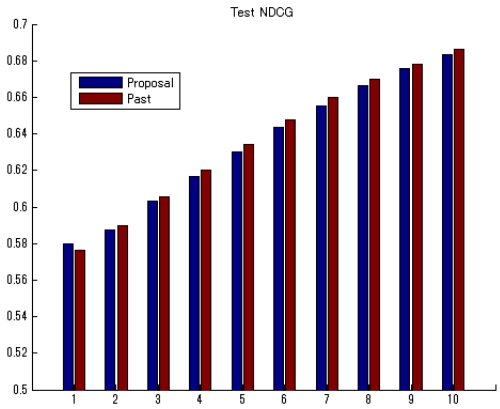

図 1: TestNDCG@1〜10 の比較

ランキング 1 位に対する学習評価値 NDCG@1 の上昇が確認され，提案法が上位を重視した学習が行えていることが確認された．また，学習時間を比べても, 従来法に比べて, 十分に早くなっていることが確認された．

7　まとめ

本研究では, RankingSVM に, 作成されるランキングにおけるドキュメントの位置情報を導入することによってランキング上位の整合性を重視した学習を行う手法を提案した．その際に，一定以上ランキングトップから離されている下位のドキュメントを学習から一時的に外すことによって学習の高速化を実現した．今後の課題は, 適切なハイパーパラメーターと初期点を効率的に定める手法の確立などが考えれる．

参考文献

[1] T-Y.Liu. Learning to Rank for Information Retrieval. *Now Publisher* 2009.

[2] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevantdocuments,. *SIGIR* pp. 41-48, 2000.

[3] C. Jui Hsieh, K. Chang, C. Lin and S. Keerthi et al. A Dual Coordinate Descent Method for Large-scale Linear SVM. *International Confer-ence on Machine Learning* , 2008

[4] 竹内 一郎, 小川 晃平, 杉山 将. 機械学習における非凸最適化問題に対するパラメトリック計画法を用いたアプローチ. *IBISML*, 2012.