Submodular Maximization with Multi-Knapsack Constraints and its Applications in Scientific Literature Recommendations

Qilian Yu[1] Easton Li Xu[2] Shuguang Cui[1]

[1]Department of Electrical and Computer Engineering
University of California, Davis

[2]Department of Electrical and Computer Engineering
Texas A&M University

2016 IEEE Global Conference on Signal and Information Processing

Dec. 7, 2016

Outline

UCDAVIS
UNIVERSITY OF CALIFORNIA

Introduction

Background

- Main Problems in Big Data Era
  - Unprecedented large datasets.
  - Heterogenous data sources.
- Submodular Optimization
  - Rich theoretical and practical features to preprocess massive data [Liu et. al. 2013].
  - Limitations on greedy fashion algorithms. [Nemhauser, Wolsey & Fisher, 1978]
- Streaming Algorithms
  - Memory required for a small portion of data.
  - Solution provided at the end of data stream.

Prerequisites

- Ground set: $V = \{1, 2, \ldots, n\}$.
- Set function: $f : 2^V \to [0, \infty)$.
- Characteristic vector: $\mathbf{x}_S = (x_{S,1}, x_{S,2}, \ldots, x_{S,n})$, where for $1 \leq j \leq n$, $x_{S,j} = 1$, if $j \in S$; $x_{S,j} = 0$, otherwise.
- Marginal gain: $\Delta_f(r|S) \triangleq f(S \cup \{r\}) - f(S)$.
  - Submodularity: $\Delta_f(r|B) \leq \Delta_f(r|A)$, for $A \subseteq B \subseteq V$ and $r \in V \setminus B$.
  - Monotone: $\Delta_f(r|S) \geq 0$, for any $S \subseteq V$ and $r \in V$.

Formulation

- Motivation: scientific literature recommendations, new recommendations, etc.
- *d*-**MASK**: Aim to **MA**ximize a monotone **S**ubmodular set function subject to a *d*-**K**napsack constraint.

$$\begin{aligned} \underset{S \subseteq V}{\text{maximize}} \quad & f(S) \\ \text{subject to} \quad & C\mathbf{x}_S \leq \mathbf{b}. \end{aligned} \tag{1}$$

  - $\mathbf{b} = (b_1, b_2, \ldots, b_d)^T$: *d*-dimension knapsack constraint vector.
  - $C = (c_{i,j})$: $c_{i,j} > 0$ is the weight of the element j with respect to the i-th knapsack resource constraint.
- *d*-MASK can be easily standardized such that $c_{i,j} \geq 1$ and $b_i = b$, for $1 \leq i \leq d$, $1 \leq j \leq n$.

	Best Performance Known Algorithms		Proposed Streaming Algorithm	
	Approx. Factor	Comput. Cost	Approx. Factor	Comput. Cost
1-Knapsack Constraint	$1-e^{-1}$ [Sviridenko, 2004]	$O(n^5)$	$1/(1+2d)-\epsilon$	$O(n \log b/\epsilon)$
d-Knapsack Constraint	$1-e^{-1}-\epsilon$ [Kulik et. al., 2009]	Polynomial		

- First to propose an efficient streaming algorithm for d-MASK, with
  - a constant-factor approximation guarantee;
  - no assumption on full access to the dataset;
  - execution of a single pass;
  - $O(b \log b)$ memory requirement;
  - $O(\log b)$ computation complexity per element;
  - only assumption on monotonicity and submodularity of the objective function.

Algorithm 1 d-KNAPSACK-STREAMING

$m := 0.$
$Q := \{[1+(1+2d)\varepsilon]^l \mid l \in \mathbb{Z}\}.$
for $v \in Q$
  $S_v := \emptyset.$
  for $j := 1$ **to** n
    for $i := 1$ **to** d
      $m := \max\{m, f(\{j\})/c_{i,j}\}.$
    end for
    $Q := \{[1+(1+2d)\varepsilon]^l \mid l \in \mathbb{Z}, \frac{m}{1+(1+2d)\epsilon} \leq [1+(1+2d)\varepsilon]^l \leq 2bm\}.$
    if $c_{i,j} \geq \frac{b}{2}$ and $\frac{f(\{j\})}{c_{i,j}} \geq \frac{2v}{b(1+2d)}$ for some $i \in [1,d]$ **then**
      $S_v := \{j\}.$
      break
    end if
    if $\sum_{l \in S \cup \{j\}} c_{i,l} \leq b$ and $\frac{\Delta_f(j|S)}{c_{i,j}} \geq \frac{2v}{b(1+2d)}$ for all $i \in [1,d]$ **then**
      $S_v := S_v \cup \{j\}.$
    end if
  end for
end for
$S := \operatorname*{argmax}_{S_v, v \in Q} f(S_v).$
return S.

Simpler Version

Algorithm 2 d-KNAPSACK-STREAMING

Initialize: Set Q.
for $v \in Q$
    for $j := 1$ **to** n
        Update Set Q.
        if j **is** *big element* **then**
            $S_v := \{j\}$.
            break.
        end if
        if j **satisfies criteria(v) then**
            $S_v := S_v \cup \{j\}$.
        end if
    end for
end for
$S := \operatorname*{argmax}_{S_v, v \in Q} f(S_v)$.
return S.

Lemma 1

Let

$$Q = \left\{ [1+(1+2d)\epsilon]^l \mid l \in \mathbb{Z}, \frac{m}{1+(1+2d)\epsilon} \le [1+(1+2d)\epsilon]^l \le 2bm \right\}$$

for some ϵ *with* $0 < \epsilon < \frac{1}{1+2d}$. *Then there exists at least some* $v \in Q$ *such that* $[1-(1+2d)\epsilon]OPT \le v \le OPT$.

Lemma 2 (Big Element)

Assume v *satisfies* $\alpha OPT \le v \le OPT$, *and there exits an element* j *such that* $c_{i,j} \ge \frac{b}{2}$ *and* $\frac{f(\{j\})}{c_{i,j}} \ge \frac{2v}{b(1+d)}$ *for some* $i \in [1, d]$.

$$f(\{j\}) \ge \frac{\alpha}{1+2d} OPT.$$

Theorem 3

Algorithm 1 has the following properties:

- *It outputs S that satisfies $f(S) \geq \left(\frac{1}{1+2d} - \epsilon\right) OPT$;*
- *It goes one pass over the dataset, stores at most $O\left(\frac{b \log b}{d\epsilon}\right)$ elements, and has $O\left(\frac{\log b}{\epsilon}\right)$ computation complexity per element.*

Theorem 4

Consider a subset $S \subseteq V$. For $1 \leq i \leq d$, let $r_{i,s} = \Delta_f(s|S)/c_{i,s}$, and $s_{i,1}, \ldots, s_{i,|V \setminus S|}$ be the sequence such that $r_{i,s_{i,1}} \geq r_{i,s_{i,2}} \geq \cdots \geq r_{i,s_{i,|V\setminus S|}}$. Let k_i be the integer such that $\sum_{j=1}^{k_i-1} c_{i,s_{i,j}} \leq b$ and $\sum_{j=1}^{k_i} c_{i,s_{i,j}} > b$. And let $\lambda_i = \left(b - \sum_{j=1}^{k_i-1} c_{i,s_{i,j}}\right) \Big/ c_{i,s_{i,k_i}}$. Then we have

$$OPT \leq f(S) + \min_{1 \leq i \leq d} \left[\sum_{j=1}^{k_i-1} \Delta_f(s_{i,j}|S) + \lambda_i \Delta_f(s_{i,k_i}|S) \right].$$

Scientific Literature Recommendations

Problem Setup

- Problem setting
  - A directed acyclic graph $G = (V, E)$ with $V = \{1, 2, \ldots, n\}$.
  - Vertex in V: an article.
  - Arc $(i, j) \in E$: paper i cites paper j.
  - A: the collection of the source papers.
- Objective
  - Select a subset S out of V to quickly detect the information spreading of A.

Problem Formulation

- Measurements
  - Length of the shortest directed path from s to a: $T(s, a)$.
  - The shortest path length from any vertex in S to a: $T(S, a) \triangleq \min_{s \in S} T(s, a)$.
  - Pre-assigned weight to each vertex $a \in A$: $W(a)$, such that $\sum_{a \in A} W(a) = 1$.
  - A given maximum penalty: $T_{\max}$.
  - The expected penalty: $\pi(S) \triangleq \sum_{a \in A} W(a) \min\{T(S, a), T_{\max}\}$.
- Formulation

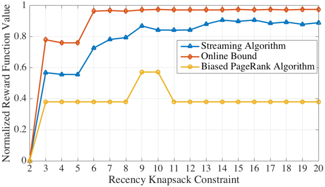
$$\begin{aligned} \underset{S \subseteq V}{\text{maximize}} \quad & R(S) \triangleq \sum_{a \in A} W(a)[T_{\max} - T(S, a)]^{+} \\ \text{subject to} \quad & C\mathbf{x}_S \leq \mathbf{b}. \end{aligned} \tag{2}$$

Experiment Setup

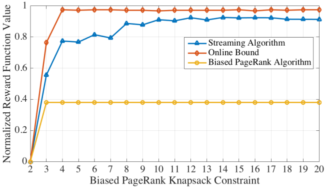
- Constraints Design
  - Recency
  - Biased PageRank Score [Gori & Pucci, 2006]
  - Reference Number
- Experiment Dataset [Joseph & Radev, 2007]
  - Over 20,000 papers in the Association of Computational Linguistics.
  - Citation network provided.

Experimental Results

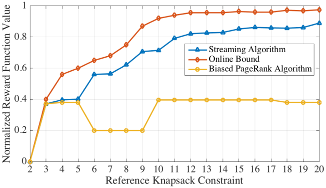
- Sensitive Analysis Setup
  - Randomly select five nodes as the source papers.
  - Set $T_{\max} = 50$ and $W(a) = 0.2$ for each source paper a.



fixed $b_2 = 10,\ b_3 = 20.$



fixed $b_1 = 20,\ b_3 = 20.$



fixed $b_1 = 20,\ b_2 = 10.$

UCDAVIS
UNIVERSITY OF CALIFORNIA

Summary

- The first streaming algorithm for d-MASK problem.
- Only a single pass through the dataset required.
- Approximation solution with a $\left(\frac{1}{1+2d} - \epsilon\right)$ factor guaranteed with much lower computation cost.
- Practical and efficient way to solve related combinatorial problem, e.g., scientific literature recommendations.

Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes,
"Submodular feature selection for high-dimensional acoustic score spaces,"
in *Proc. 2013 IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, BC, May 2013, pp. 7184–7188.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher,
"An analysis of approximations for maximizing submodular set functions–I,"
Math. Program., vol. 14, no. 1, pp. 265–294, Dec. 1978.

M. Sviridenko,
"A note on maximizing a submodular set function subject to a knapsack constraint,"
Oper. Res. Lett, vol. 32, pp. 41–43, Jan. 2004.

A. Kulik, H. Shachnai, and T. Tamir,
"Maximizing submodular set functions subject to multiple linear constraints,"
in *Proc. 20th Annu. ACM-SIAM Symp. Discrete Algor.*, New York, NY, Jan. 2009, pp. 545–554.

M. Gori and A. Pucci,
"Research paper recommender systems: A random-walk based approach,"
in *Proc. 2006 IEEE/WIC/ACM Int. Conf. Web Intell.*, Hong Kong, Dec. 2006, pp. 778–781.

M. T. Joseph and D. R. Radev,
"Citation analysis, centrality, and the ACL anthology,"
Tech. Rep., CSE-TR-535-07, Oct. 2007.