Distributed Maximization of “Submodular plus Diversity” Functions for Multi-label Feature Selection on Huge Datasets

Mehrdad Ghadiri and Mark Schmidt
Dept. of Computer Science, University of British Columbia

Overview

Multi-label feature selection problem is in need of scalable methods because of the rapid growth of the size of datasets.

Here we develop a theoretical modeling for this problem. We formulate it as a “submodular plus diversity” optimization problem and show that an approximation algorithm can be used to maximize this optimization problem in a distributed setting.

Main Contributions:

- Formulating the multi-label feature selection problem as a combinatorial optimization problem. Namely, as the maximization of the sum of a monotone submodular function and a sum-sum diversity function.
- Presenting a greedy algorithm for such a combinatorial optimization problem in the distributed and streaming settings and showing it achieves a constant factor approximation.
- Performing an empirical study on the resulted multi-label feature selection method and comparing it to the state-of-the-art centralized feature selection methods.

Distributed Multi-label Feature Selection

Samples

Features									Labels		
5	6	7	2	6	3	5	8	4	3	4	6
5	8	2	4	5	9	2	1	3	5	8	12
6	4	8	9	4	3	4	3	1	4	18	27
5	3	2	5	8	4	2	1	5	8	10	15

Properties of the dataset

- A small number of samples
- A huge number of features

Therefore we need a **Filter** method with **Vertical Distribution** of data

A Theoretical Modeling

Goal: select **non-redundant relevant** features

- Model the **dis-similarity** of features with a **metric** distance function.
- Model the **relevance** of features with a **submodular** function.

Set of features: $U = \{u_1, \dots, u_n\}$
Set of labels: $L = \{\ell_1, \dots, \ell_t\}$

The following is a dis-similarity/distance measure between pairs of features.
Metric distance (Normalized Variation of Information):

$$d(u_i, u_j) = 1 - \frac{I(u_i, u_j)}{H(u_i, u_j)}$$

The following is a submodular function that represents the relevance of a subset of features to the set of labels.
Submodular function ($S \quad P$):

$$g(S) = \sum_{\ell \quad L} \mathrm{top}^{\mathrm{p}}_{u \quad S}\{MI(u, \ell)\}$$

H is the joint entropy, I is the mutual information, MI is the normalized mutual information, and $\mathrm{top}^{\mathrm{p}}$ is the sum of the p largest number in the associated set.

The $\mathrm{top}^{\mathrm{p}}$ function causes the formulation to select at least p relevant features for each label. In the extreme cases of $p = 1$ and $p = n$, one or few features can dominate the formulation and prevent it to find a good set of features.

Optimization Problem: Maximize the following subject to $|S| \leq k$

Redundancy	Relevance
Diversity function	Submodular function

$$\lambda \sum_{u_i, u_j \quad S} d(u_i, u_j) + (1 - \lambda)\frac{k(k-1)}{2p|L|} g(S)$$

$\lambda \quad [0, 1]$ is a hyper-parameter.

Theoretical Problem

“Submodular plus Diversity”

Maximize the sum of a diversity function and a monotone submodular function subject to a cardinality constraint.

Theoretical Contribution

Maximizing a **“submodular plus diversity”** function in **distributed** and **streaming** settings within a constant-factor approximation.

Algorithms

Algorithm 1: Greedy

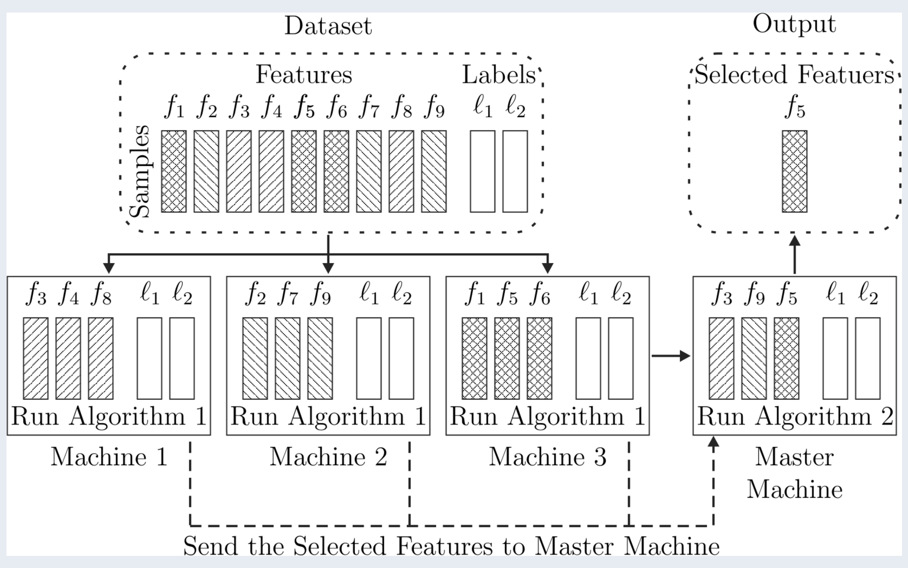
Input: Set of features U, set of labels L, number of features we want to select k.
Output: Set $S \quad U$ with $|S| = k$.
$S \leftarrow \{\arg\max_{u \quad U} g(\{u\})\}$;
forall $2 \leq i \leq k$ **do**
  $u \quad \leftarrow \arg\max_{u \quad U \setminus S} \; g(S \quad \{u\}) - g(S) + \sum_{v \quad S} d(v, u)$;
  Add $u \quad$ to S;
Return S;

Algorithm 2: AltGreedy

Input: Set of features U, set of labels L, number of features we want to select k.
Output: Set $S \quad U$ with $|S| = k$.
$S \leftarrow \{\arg\max_{u \quad U} g(\{u\})\}$;
forall $2 \leq i \leq k$ **do**
  $u \quad \leftarrow \arg\max_{u \quad U \setminus S} \; \frac{1}{2}(g(S \quad \{u\}) - g(S)) + \sum_{v \quad S} d(v, u)$;
  Add $u \quad$ to S;
Return S;

Composable Core-sets

Distributed Setting



Streaming Setting

In the streaming setting, we have a machine that receives the data from a random stream. Therefore, it can pick chunks of data and do the Greedy algorithm on them and store the selected features in the memory. Then when the stream is ended, it produces the final set of features by running the AltGreedy algorithm on the stored features.

A **composable core-set** returns subsets which their **union** contains an **approximate solution**.

Here we use **Randomized Composable Core-sets** which means the data is randomly partitioned.

Theoretical Result

We show that a randomized composable core-set finds a $\frac{1}{31}$**-approximate solution** in expectation.

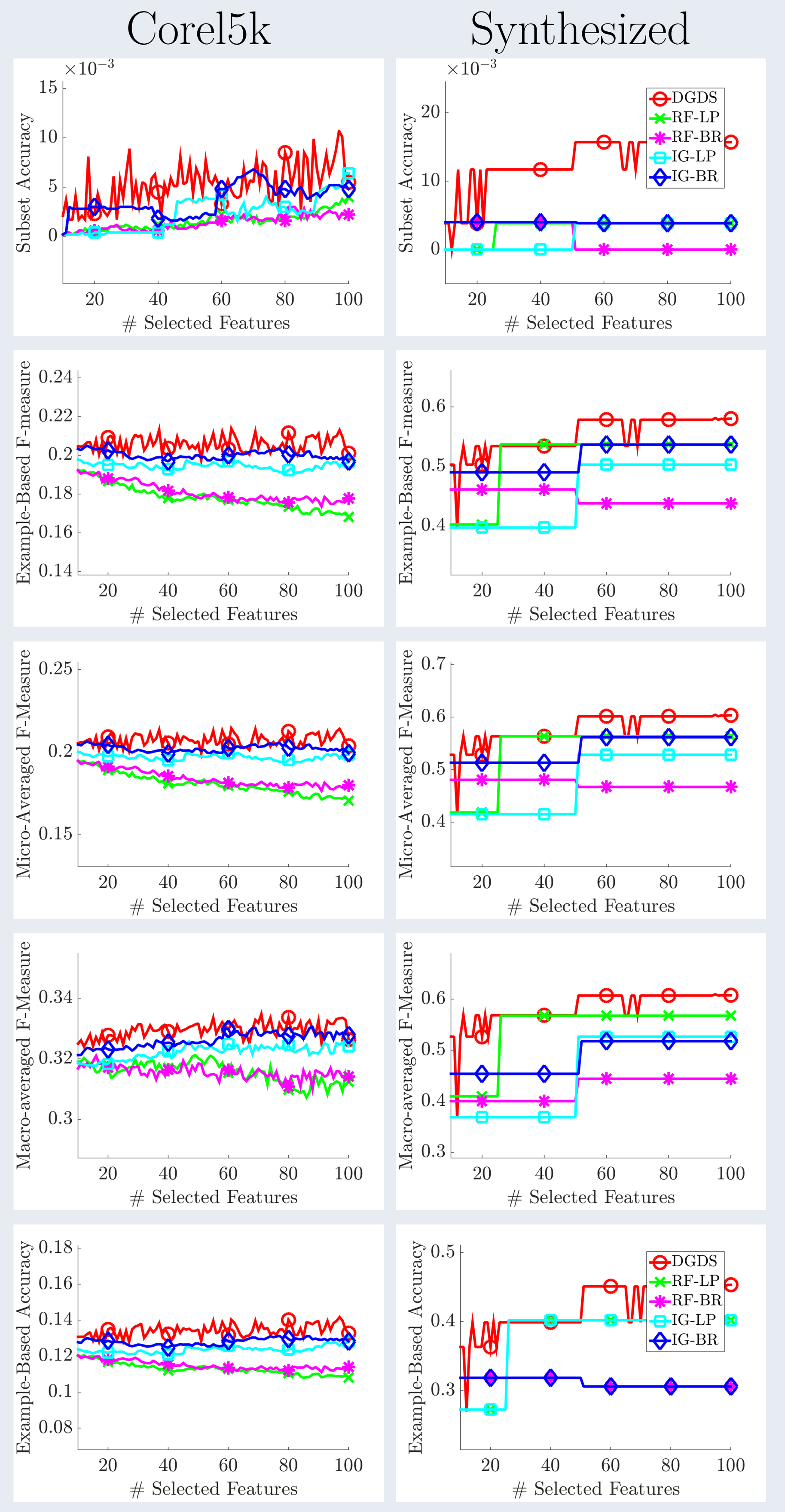
Proof idea: Proof relies on the notion of β-niceness of an algorithm defined by Mirrokni et al (STOC’15). An algorithm is **β-nice** if the marginal gain of adding an element to its output is less than β times the average contribution of the elements of the output.
This property shows that the output of an algorithm is “good enough” in the sense that adding other elements to its output does not increase the objective too much. This provides a theoretical bound for the optimum solution.
We show that the greedy algorithm is **5-nice** for this class of functions and using this, we conclude our result.

Empirical Results

Speed-up

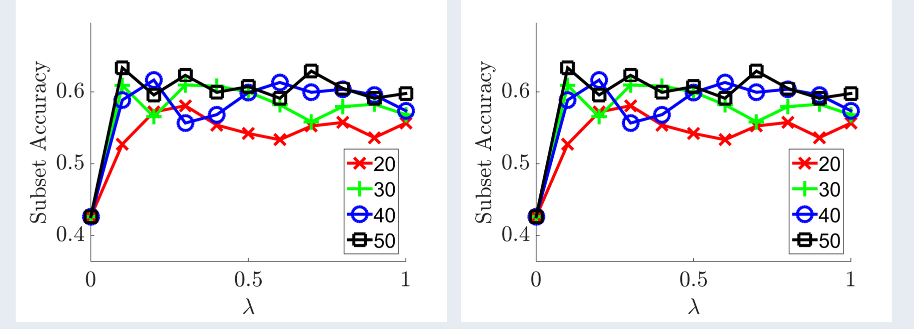
Dataset Name	# Selected Features	# Machines	Distributed Algorithm Runtime	Centralized Algorithm Runtime	Speed-up
RCV1V2	10	69	2.8m	1h 33m	33.2
	50	31	10.8m	2h 30.0m	15.1
	100	22	20.3m	3h 39m	10.8
	200	16	47.0m	6h 16.8m	8.0
TMC2007	10	71	4.6m	2h 32.5m	33.4
	50	32	24.2m	6h 24.7m	15.9
	100	23	59.5m	11h 6.2m	11.2
	200	16	2h 41.3m	20h 49.8m	7.7

Feature Selection Performance Compared to State-of-the-art Centralized Methods



- Label powerset (LP) and binary relevance (BR) convert a multi-label dataset to one or multiple single-label datasets.
- ReliefF (RF) and information gain (IG) are two methods for single-label feature selection.

Effect of λ



The legend values indicate the number of selected features.

Related Work

- Borodin et al (PODS’12) show a half approximation for maximizing a “submodular plus diversity” function in the centralized setting.
- Abbasi-zadeh et al (AAAI’17) show a quarter approximation for maximizing a diversity function in a distributed setting. They use this framework for single-label feature selection.
- Mirrokni et al (STOC’15) show a 0.27-approximation for maximizing a submodular function in a distributed setting.
- Dasgupta et al (ACL’13) consider the maximization of the sum of a submodular function and other diversity functions (sum-sum diversity, minimum spanning tree, and minimum distance).