Graph-based Semi-Supervised Learning for Text Categorization through Supervised Random Walks

Tom Shen, advised by William W. Cohen

Background

Semi-supervised learning (**SSL**) is a core task in machine learning. Given a large amount of unlabeled data, and a small amount of labeled data, we want to assign labels to the unlabeled data.

We can often represent the data as a graph, where nodes represent data instances, and edges between nodes are weighted by some similarity metric. We then perform **label propagation** to propagate the labels from the instances with pre-assigned labels (**seed** instances) to unlabeled instances.

Given a set of text documents (e.g. webpages, news articles), we often want to categorize them. This is called **text classification**.

We can build a graph representation of a set of documents by connecting documents to words contained in those documents. Humans can label a small portion of the documents, and then we can perform graph-based SSL to propagate those labels to the rest of the documents.

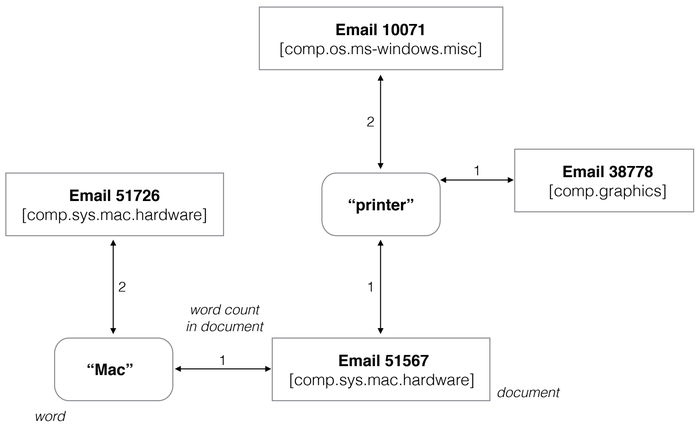

Figure 1. A graph representation of a subset of the 20 Newsgroups dataset

Programming with personalized PageRank (**ProPPR**) is a first-order probabilistic language for logical inference. Traditionally, queries are answered through *grounding*, which involves mapping queries to propositional representations, and then performing inference.

ProPPR uses an implementation of **supervised random walks** (**SRW**), which were first developed as a method to predict links between nodes in a social network graph. Given a graph with feature-vector annotated edges, a *personalized PageRank* process is performed through optimization of a pairwise ranking loss function through *stochastic gradient descent,* for speed and easy parallelization.

Motivation

We want to see if ProPPR's implementation of SRW can be used on its own for label propagation. In particular, we want to see how it compares to state-of-the-art label propagation methods.

Algorithm

Input: Undirected, weighted graph $G = (V, E)$, mapping of seed nodes to labels S, list of labels L
Output: Mapping of nodes to ranked list of labels R
Let $V' = V$, $E = []$;
Create directed, feature-vector weighted graph $G' = (V', E')$;
Add start node s to V';
for $(u, v, w) \in E$ **do**
  Add (u, v) with feature vector $[w, 0]$ to E';
  Add (v, u) with feature vector $[w, 0]$ to E';
end
Let $R = \{v \mapsto [] \mid v \in V\}$;
for $l \in L$ **do**
  Let $G'' = (V'', E'')$ be a copy of G';
  for $(v \mapsto l') \in S$ **do**
    if $l = l'$ **then**
      Add (s, v) with feature vector $[0, 1]$ to E'';
    end
  end
  Run SRW on G'', and let M be the output;
  for $v \in V$ **do**
    Append $(M[v], l)$ to $R[v]$;
  end
end
for $v \in V$ **do**
  Sort in descending order $R[v]$;
end

Figure 2. Pseudocode for the algorithm

We generate a graph for each label. This graph is the document-word representation of the instances, with the addition of a start node for the random walk. This start node is connected to the seed instances associated with that label. After running SRW on the graph, we have a weight associated with each instance. We can then generate a label ranking for each instance based on its weight in the associated label's graph.

Existing Algorithms

Gaussian random fields (**GRF**) is one of the earliest label propagation methods. It "prefers" smooth labelings, since the score of the label for a node is the weighted average of the scores for that label of the node's neighbors. Additionally, label assignments where two nodes connected by a highly weighted edge but are assigned different labels are penalized.

Modified adsorption (**MAD**) is a method based on *adsorption*. Adsorption is an algorithm that involves a random walk over a graph where, at each vertex *v*, we either stop the walk and return a predefined vector Y_v, continue the walk to a randomly chosen neighbor of *v* based on edge weights, or abandon the walk and return a zero vector. MAD retains most of the properties of the adsorption algorithm, but with a well-defined objective function to minimize.

Datasets

A commonly used text classification dataset is **20 Newsgroups** (**20NG**). It is a collection of approximately 20 thousand emails, evenly spread out between 20 different mailing lists. These mailing lists span a variety of topics: automobiles, politics, religion, science, sports, and technology.

Datasets (continued)

WebKB is a set of approximately four thousand webpages collected by the World Wide Knowledge Base project from computer science departments of four universities (Cornell, Texas, Washington, and Wisconsin) in 1997. We are using a subset of the webpages with four categories: student, faculty, course, and project.

Results

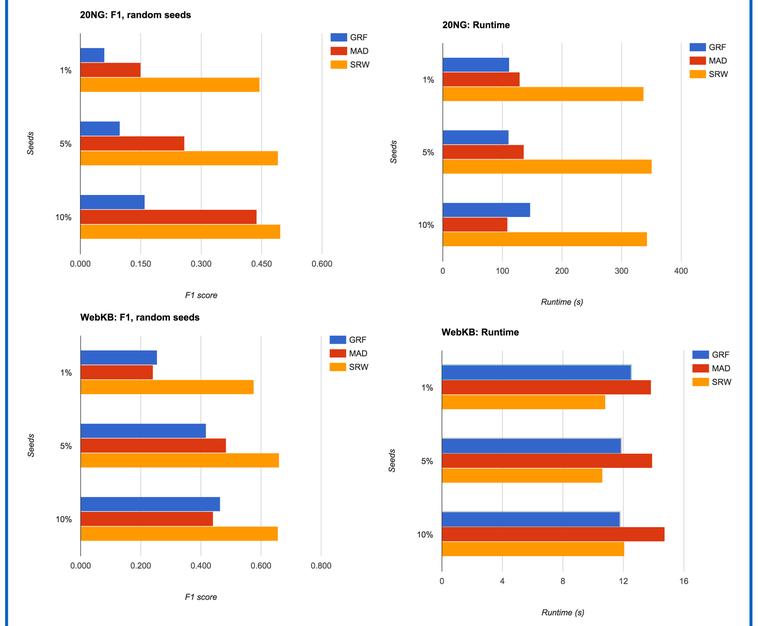

Figure 3. Plots of F1 score and runtime on 20 Newsgroups

We compared the performance of SRW to GRF and MAD on the two text classification datasets. SRW outperformed both GRF and MAD on both datasets. It had greater runtime on 20 Newsgroups, but comparable runtime on WebKB.

However, with parallelization, SRW attains up to a 5x speedup, which makes its runtime better than the other algorithms.

Thus, SRW has been shown to be an effective graph-based SSL algorithm.

Future Work

The runtime of supervised random walks is directly proportional to number of classes, since it runs the algorithm on a graph generated for each class. Since the same is not true for Gaussian random fields and modified adsorption, we could investigate adapting parts of those algorithms to improve performance of SRW on datasets with many classes.

Additionally, speedup from parallelization falls off at a certain point, so we could investigate reducing communication costs and sequential work to improve parallel performance.