Project Proposal for CS 4780 Machine Learning

Cornell University

1. **The Team:**
Michael Flashman (mtf53@cornell.edu),
Andrey Gushchin (avg36@cornell.edu),
Hyung Joo Park (hp255@cornell.edu).
2. **Motivation**:
Considering the pace at which scientific papers are being written and published, it is becoming more and more time consuming to individually read each paper to get a good sense of its overall quality. It would be interesting to see if effective algorithms which can help facilitate the process of discerning what constitutes a good or bad paper can be devised.
3. **Statement of the Problem/Task**:
- Using the number of citations/year as a metric of quality, is it possible to predict this value based on the relative frequency of words that appear in the paper?
- Is it possible to predict the quality of a paper based on its metadata (author's name, his/her institution, submission date, etc.)?
- Is prediction based on the relative frequency of words better than the one based on the metadata?
- Which features are more relevant for this problem (feature selection problem)?
- Would it be possible to predict for how long a paper may stay relevant?
4. **General Approach:**
We will start with a two-class classification problem and non-kernel SVM. For this we will binarize the articles labels (citation counts); for example, label y_i is 0, if number of citations for article i is less than 10; it is 1, otherwise. First, we will use word frequencies as the features, and then articles metadata. After that we will combine these two types of features together and will see which of the features are the most relevant. Our next step will be to apply kernel SVM and find the best type of a kernel for this problem. Then we can use some other algorithms and compare their performance with the result of SVM. We are also planning to solve a regression-SVM to obtain numeric predictions of articles citation counts. We are going to compare our results with the results of similar projects (see references).
5. **Resources:**
We will perform our citation-count learning task on 27770 high energy particle physics theory (HEP-TH) articles published to the arXiv e-print database during the period from January 1993 to April 2003.
Citations counts for each article will be obtained by constructing the article citation network and adding up relevant citations. The citation network (HEP-TH citation network data set) is provided by of the Stanford Network Analysis Project, and can be found at http://snap.stanford.edu/data/cit-HepTh.html. Citations to and from papers outside the article collection are ignored. The resulting network contains a total of 352807 citations.

We will train our learning algorithm on features drawn from two different data sets. The first data set consists of bag-of-word representations for the full text of each article. Specifically, each article is represented as the collection of n-grams that appear in the full text of the article together with n-gram frequencies. Here, the full text of an article includes title, authors, affiliation, body and references. Rather than specify a fixed n, our bag-of-word model includes 1-grams, 2-grams, and 3-grams. Exactly which words are left as 1-grams, and which words are grouped together as 2-grams and 3-grams is determined on the statistical significance of the grouping across the full collection of articles. The complete bag-of-word model for the collection of articles contains over 2.5 million distinct n-grams.

This data set is provided generously by Paul Ginsparg (as part of ongoing work with Michael Flashman).

Our second data set consists of article metadata, and is provided as part of the HEP-TH citation network data set from above. This metadata includes such information as article title, authors, abstract, journal ref, and submission date.

Support vector machines used for our learning task will be handled by Thorston Joachims' *SVM light* software.

6. **Schedule:**

By 11/5: prepare the data, translate it to *SVM light* format.

By 11/12: with non-kernel SVM perform two-class classification using frequencies of words in the articles as the features, then using metadata of the articles.

By 11/19: with non-kernel SVM perform two-class classification using both frequencies of words in the articles and the metadata.

By 11/26: solve the same problem using SVM with kernels. Try to find the best kernel for this problem. Apply several other algorithms (kNN, Neural Networks, Naïve Bayesian algorithm) for this problem.

By 12/03: solve regression problem using SVM (predict number of citations for articles), obtain and discuss the results.

7. **References:**

1) Acuna, D. E.; Allesina, S. & Kording, K. P. Future impact: Predicting scientific success
Nature, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., **2012**, *489*, 201-202
2) Adler, R.; Ewing, J. & Taylor, P.Citation statistics
Statistical Science, **2009**, *24*, 1
3) Gehrke, J.; Ginsparg, P. & Kleinberg, J. M. Overview of the 2003 KDD Cup
SIGKDD Explorations, **2003**, *5*, 149-151
4) Ibáñez, A.; Larrañaga, P. & Bielza, C. Predicting citation count of Bioinformatics papers within four years of publication
Bioinformatics, **2009**, *25*, 3303-3309
5) Livne, A.; Adar, E.; Teevan, J. & Dumais, S. Predicting Citation Counts Using Text and Graph Mining
iConference, **2013**
6) Manjunatha, J. N.; Sivaramakrishnan, K. R.; Pandey, R. K. & Murthy, M. N. Citation Prediction Using Time Series Approach KDD Cup 2003 (Task 1)
SIGKDD Explor. Newsl., ACM, **2003**, *5*, 152-153
7) McNamara, D.; Wong, P.; Christen, P. & Ng, K. Li, J.; Cao, L.; Wang, C.; Tan, K.; Liu, B.; Pei, J. & Tseng, V. *(Eds.)* Predicting High Impact Academic Papers Using Citation Network Features
Trends and Applications in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, **2013**, *7867*, 14-25
8) Perlich, C.; Provost, F. & Macskassy, S. Predicting citation rates for physics papers: constructing features for an ordered probit model
SIGKDD Explor. Newsl., ACM, **2003**, *5*, 154-155
9) Yan, R.; Tang, J.; Liu, X.; Shan, D. & Li, X. Citation count prediction: learning to estimate future citations for literature
Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, **2011**, 1247-1252