Prediction of Research Impact : A Case Study for Nanotechnology

Raisul Islam (raisul@stanford.edu), Patrick Tae (ptae@stanford.edu)

Overview

The goal of the this project is to predict the impact of a research paper/project using keywords related to the paper's title, which could help give researchers a feel for the level of interest in a proposed project. An effective measure of research impact is the number of citations received over the years, so our proposed algorithms attempt to predict the number of citations per year a paper will get in the area of nanotechnology, based on its title keywords.

Dataset / Data Processing

Top papers on Google Scholar from 3 journals:

- Nanoletters, ACS Nano, Nature Nanotechnology
- 523 papers total containing titles, publication years, and citation counts

Input raw data:

Title: {**Fracture of Silicon Nanoparticles During Lithiation**} -> x
Citations: **# of citations / Years passed since it appeared**->y

↓

-Remove non-alphanumeric characters, replace with spaces.
-Make all characters lowercase, split on whitespaces to form a vector of keywords.
'fracture' 'of' 'silicon' 'nanoparticles' 'during' 'lithiation'

↓

-Remove prepositions/small unimportant words
'fracture' 'silicon' 'nanoparticles' 'during' 'lithiation'

↓

-Find root substring of each word by comparing to the vocabulary. If it is a new word, it is added to the vocabulary.
'fracture' 'silicon' 'particle' 'during' 'lithiation'

Theory / Algorithms

Gaussian Discriminant Analysis

Divide the dataset into b bins based on number of citations/year. Calculate $p(x=b|y)$ and $p(y)$ from the dataset. Then make prediction using max probability given by:

$$p(y=b|x)=\frac{p(x|y=b)p(y=b)}{\sum_{j=1}^{k}p(x|y=j)p(y=j)}$$

Softmax Regression

Here we do a linear regression to fit the entire space of vocabulary words to predict a given title.

$$h_{\theta_l}(x^{(i)})=\phi_l$$

$$p(y=l|x^{(i)};\theta)=\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}=\phi_l$$

$$\theta_l := \theta_l+\alpha y^{(i)}\left(1-h_{\theta_l}(x^{(i)})\right)x^{(i)}$$

K-means Clustering

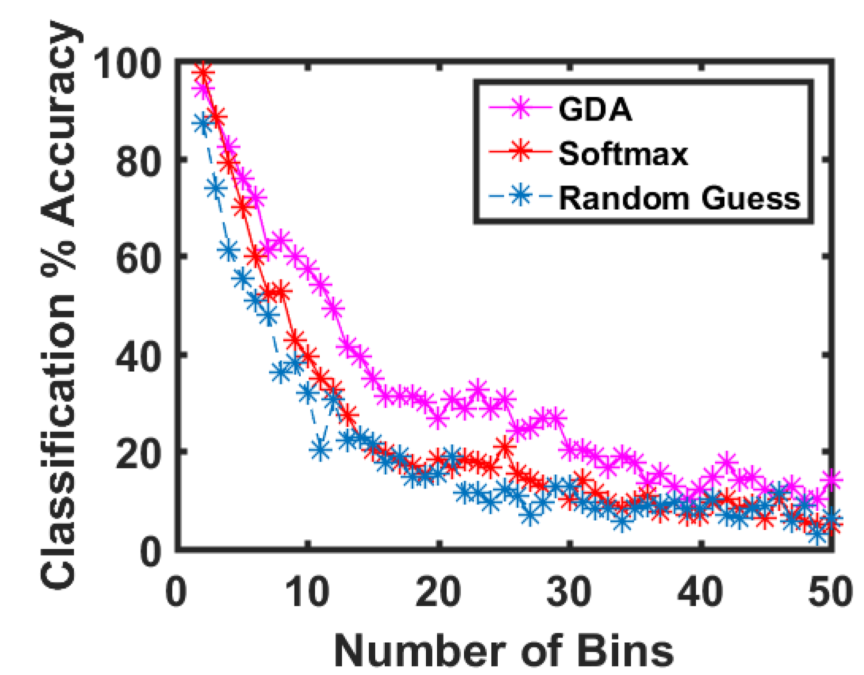
Cluster the titles based on similarity, i.e. number of common words.

Algorithm:

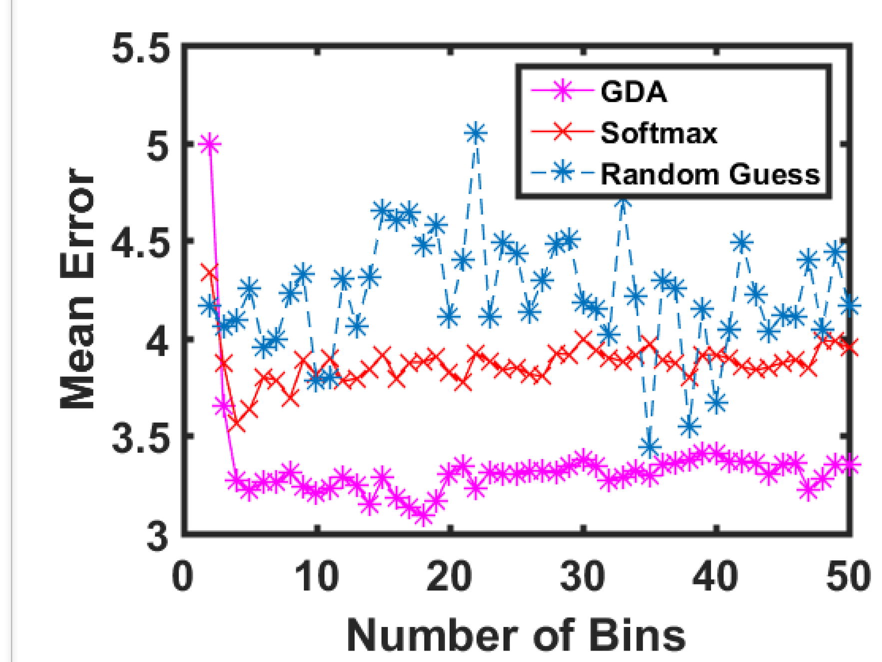
1. Sort titles into clusters based on similarity to the cluster keywords
2. Redefine the each cluster's keywords to be the most common words in the cluster.

Results

Classification Accuracy



Mean Error



Error metric:

$$E=\frac{\left\|C_{true}=c_{y=b}\right\|_2}{m}$$

$c_{y=b}$ is the center of the predicted bin.

Training / Test Error

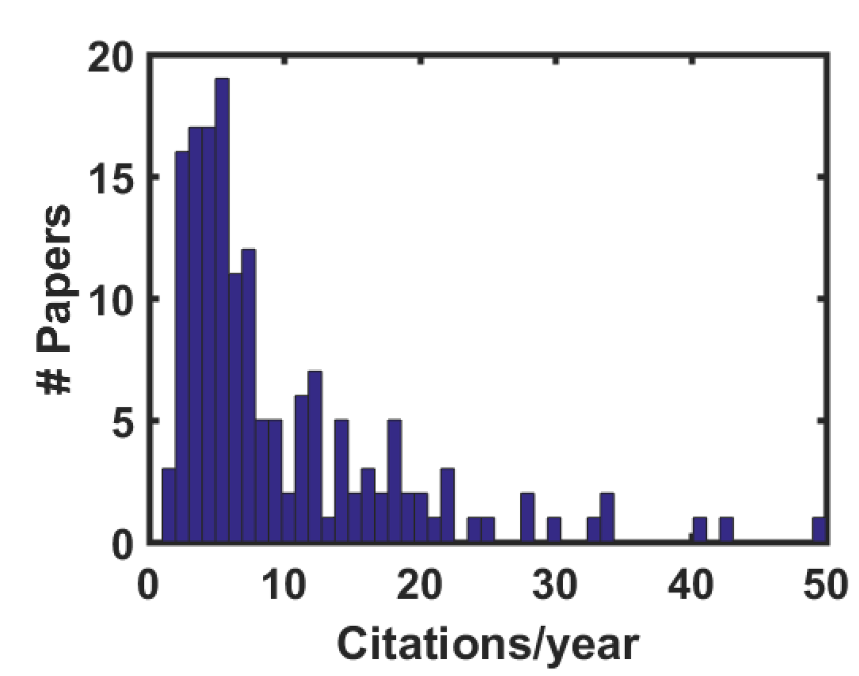
Algorithm	Train Error	Test Error
GDA	1.81	3.31
Softmax	0.996	3.83
Random Guess	4.28	4.28

Random guessing done by randomly choosing a bin over the distribution of binned citation data.

Clusters

High Performance	Solar cells	Transistor Technology	Battery Technology	Drugs/ Medical	General Materials
graphene	solar	layer	graphene	plasmon	electrode
high	cells	mos2	particle	surface	ultra
carbon	graphene	transistor	cation	sensor	mos2
performance	layer	effect	dimensional	particle	structure
super	oxide	particle	based	nanorod	material
nanotube	electron	silicon	material	therapy	crystal
structure	sensitized	field	micro	inter	super
hybrid	perovskite	single	cells	cancer	sheet

Data Distribution



Discussion / Conclusion

- We observe GDA to have lower test error than softmax, but each offer only a modest improvement over random guessing.
- Title words alone are insufficient information for a meaningful prediction. In future should consider abstract, authors, references, and even body text.
- But clustering can successfully group titles into current relevant research categories.