Important Citations in Paper

Mentor: Prof. Purushottam Kar
By: Decoders
Department of Computer Science and Engineering
IIT Kanpur

Outline

1. Abstract
2. Related Work
3. Data Generation and Extraction
4. Methods Used
5. References

Abstract

- Problem of assigning most suitable reviewers to papers in conferences.
- Currently used TPMS uses a somewhat bag-of-words approach to compare papers, doesn't exploit content properties.
- The aim of this project is to attempt to use content and context of the citation while deciding best reviewers.
- We remodel the problem into finding top citations for each paper, which in-turn can be used while deciding reviewers.

Related Work

- Toronto Paper Matching System
- Identifying Meaningful Citations
- MyReview

Data Generation and Extraction

- We used papers from ACL anthology, having dataset of 465 annoted pairs (cited paper, citing paper). It was categorized as citation types of related work, comparison, using the work and extending the work with labels 1,2,3,4 respectively
- We changed the data to only two labels 0 and 1
- The dataset of papers behind these citations are scraped using local scripts.
- To extract the text from the PDF files we used 'pdftotext2'
- The text in papers is normalized by removing diacritics with python script
- In order to segment the paper into sections we used ParsCit

Selected Features

1. **The Number of Direct Citations**
   This feature counts the total number of direct citations.
2. **Number of direct citations per section**
   This feature counts the number of direct citations corresponding to each section.
3. **1/(Number of References)**
   This is the inverse of the length of the papers citing list.
4. **Similarity between Abstracts**
   This is the tf-idf score between abstracts.
5. **PageRank**
   This is the PageRank score between each paper and its reference.
6. **Importance of citation derived from sentence**
   Vectorize the words and learn independent classifier to get importance of citation.

Importance of citation derived from sentence

1. **Getting the sentences** We wrote a script which yields sentences corresponding to each citation present in the paper.
2. **Creating the Dataset** We annotated around 900 such sentences containing citations for this purpose. A class 0 or 1 was awarded based on annotator's deduction about usefulness of the citation as visible from the paper.
3. **Training** We then used several techniques for training and validation. The best classifier based on highest validation score was selected.

Classifier	Accuracy(%)	Precision(0)	Precision(1)	Recall(0)	Recall(1)
RBF SVM	77	0.89	0.39	0.82	0.53
Bernoulli Naive Bayes	82	0.84	0.5	0.97	0.16
Multinomial Naive Bayes	82	0.83	0	0.99	0
Gaussian Naive Bayes	71	0.85	0.26	0.79	0.34
Linear SVM	84	0.85	0.75	0.99	0.19
KNN(k=5)	81	0.83	0.33	0.97	0.06

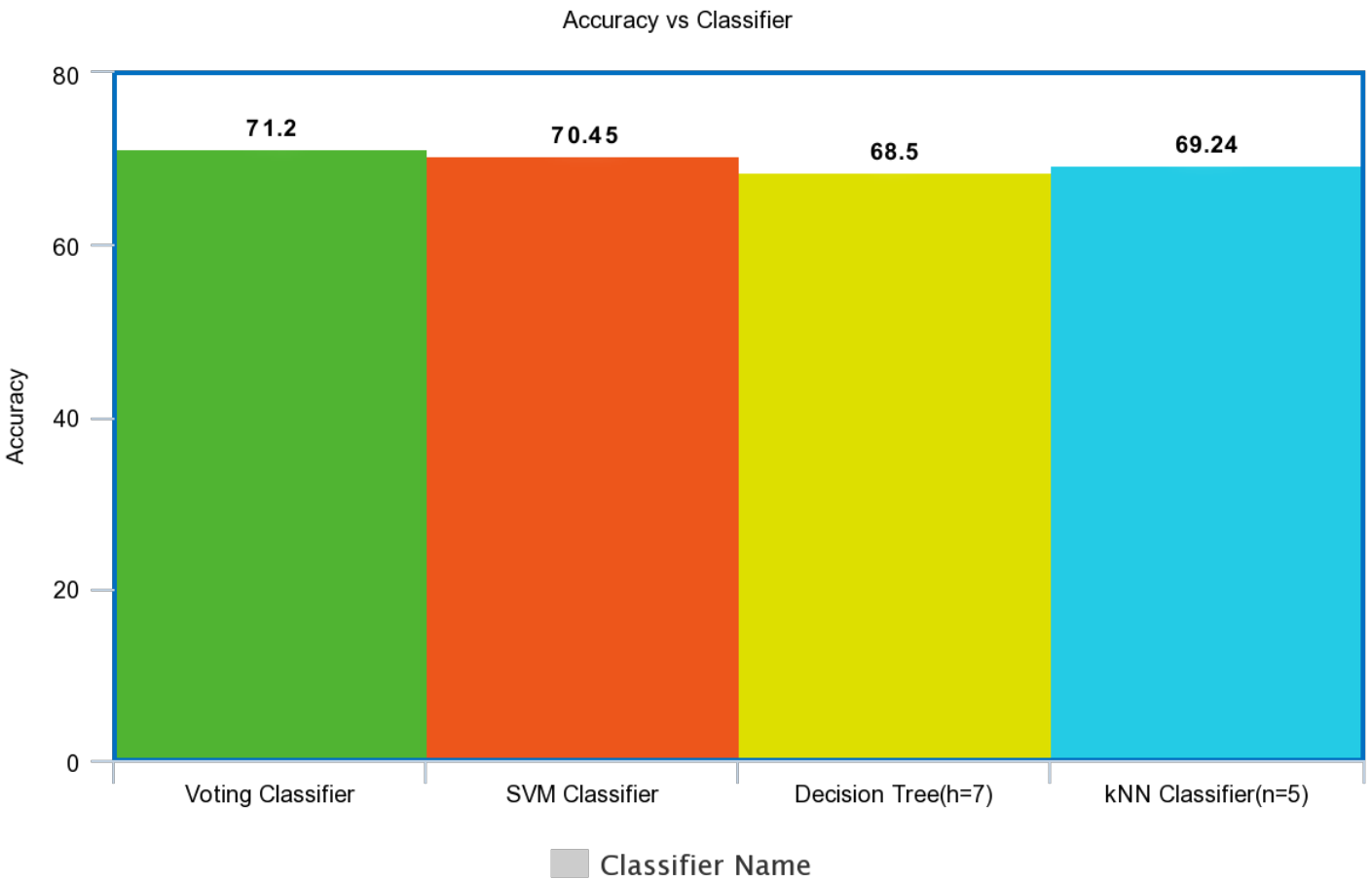
Other features tried

1. **Citation in table or caption**
   This computes the citations that appear in a table or caption.
2. **Author Overlap**
   This return true if the author has cited himself/herself in the citation.
3. **Field of the cited paper**
   We tried to learn the topic on which the paper was based by using NMF (Negative Matrix Factorization) and 20 topics as hyperparameter.

Algorithms Used

- We have used sklearn's SVM(rbf kernel), Decision Tree(height=10) and KNN(neighbors=5).
- We have also used sklearn's VotingClassifier with base classifiers Decision Tree(depth=4), KNN(number of neighbors=5) and SVM with rbf kernel.

Accuracy vs Classifier



Feature Importance according to Decision Tree Classifier

Feature Name	Importance
Author overlap	0.01
1 /# of references	0.06
Similarity between abstracts	0.17
PageRank	0.29
# direct citations	0.07
# citations in experiment sections	0.05
# references/total_references	0.20

References

- (TPMS) http://www.cs.toronto.edu/ lcharlin/papers/tpms.pdf
- Identifying meaningful citations (https://iths.pure.elsevier.com/en/publications/identifying-meaningful-citations)