parameters

Gerbrich Ferdinands

1/14/2020

ASReview takes the following parameters/arguments:

- a model
- a query strategy
- a balance strategy (fixed)
- a feature extraction strategy
- number of training data

The goal: Use these inputs to predict relevance of papers.

Numerical representation of texts To perform a learning algorithm on the abstract, the content of the abstracts needs to be represented numerical. To perform learning algorithms on the abstracts, their content has to be represented in a numerical way. This is done by transforming the textual content of the abstracts into numerical feature vectors. A classical example representation of texts is by 'bag of words' features, where for each text the number of occurrences of each word is stored. This leads to n features, where n is the number of distinct words in the texts. (Pedregosa et al. 2011)

ASReview implements several feature extraction strategies. The following will be compared:

- Doc2Vec (Le and Mikolov 2014)
- Tfidf
- SBERT
- embeddingIdf

The model is typically a learning algorithm used to predict the relevance of text. The following models will be compared:

- Naive Bayes
- Random Forests
- Support Vector Machine
- Logistic Regression
- **Dense Neural Network**

Active learning = increasing classification performance with every query. The query strategy determines the way unlabeled papers are queried to the researcher.

(Danka and Horvath, n.d.)

The balance strategy

	Configurations
Models	Naive Bayes, Random Forest, Support Vector Machine, Logistic Regression
Query Strategies	Cluster Sampling, Maximum Sampling, Cluster * Maximum Sampling, Maximum * Uncertainty Sampling, Maximum * Random Sampling, Cluster * Uncertainty Sampling, Cluster * Random Sampling
Feature extraction strategies	Doc2Vec, tf-idf, sbert, embeddingIdf
Training data [included/excluded]	10/10, 5/5, 5/10

Models

Naive Bayes

Naive Bayes assumes all features are independent given the class value. (Zhang 2004)

ASReview uses the `MultinomialNB` from the scikit-learn package (Pedregosa et al. 2011), that implements the naive Bayes algorithm for multinomially distributed data. `nb`

Hyperparameters * alpha - accounts for features not present in learning samples and prevents zero probabilities in further computations. ?

Random Forests

Support Vector Machine

Logistic Regression

Dense Neural Network

Query Strategies

- Max - Choose the most likely samples to be included according to the model
- Uncertainty - choose the most uncertain samples according to the model (i.e. closest to 0.5 probability) (Lewis and Catlett 1994)
- Random - randomly selects abstracts with no regard to model assigned probabilities.
- Cluster - Use clustering after feature extraction on the dataset. Then the highest probabilities within random clusters are sampled

The following combinations are simulated:

- cluster
- max
- cluster * random
- cluster * uncertainty
- max * cluster
- max * random
- max * uncertainty

Balance Strategies

Feature Extraction Strategies

Combinations

This leads to 273 combinations of configurations.

- Naive bayes only goes with tfidf feature extraction.
- For the feature extraction strategies we will focus on doc2vec and tfidf. (but will compute all 4)
- This leads to 3 * 7 * 4 * 3 + 1 * 7 * 1 * 3 = 273 combinations.

References

Appendix

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
nb	cluster	tfidf	10/10
nb	max	tfidf	10/10
nb	max * cluster	tfidf	10/10
nb	max * uncertainty	tfidf	10/10
nb	max * random	tfidf	10/10
nb	cluster * uncertainty	tfidf	10/10
nb	cluster * random	tfidf	10/10
nb	cluster	tfidf	5/5
nb	max	tfidf	5/5
nb	max * cluster	tfidf	5/5
nb	max * uncertainty	tfidf	5/5
nb	max * random	tfidf	5/5
nb	cluster * uncertainty	tfidf	5/5
nb	cluster * random	tfidf	5/5
nb	cluster	tfidf	5/10
nb	max	tfidf	5/10
nb	max * cluster	tfidf	5/10
nb	max * uncertainty	tfidf	5/10
nb	max * random	tfidf	5/10
nb	cluster * uncertainty	tfidf	5/10
nb	cluster * random	tfidf	5/10
rf	cluster	doc2vec	10/10
rf	max	doc2vec	10/10
rf	max * cluster	doc2vec	10/10
rf	max * uncertainty	doc2vec	10/10
rf	max * random	doc2vec	10/10
rf	cluster * uncertainty	doc2vec	10/10
rf	cluster * random	doc2vec	10/10
rf	cluster	doc2vec	5/5
rf	max	doc2vec	5/5

(continued)

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
rf	max * cluster	doc2vec	5/5
rf	max * uncertainty	doc2vec	5/5
rf	max * random	doc2vec	5/5
rf	cluster * uncertainty	doc2vec	5/5
rf	cluster * random	doc2vec	5/5
rf	cluster	doc2vec	5/10
rf	max	doc2vec	5/10
rf	max * cluster	doc2vec	5/10
rf	max * uncertainty	doc2vec	5/10
rf	max * random	doc2vec	5/10
rf	cluster * uncertainty	doc2vec	5/10
rf	cluster * random	doc2vec	5/10
rf	cluster	tfidf	10/10
rf	max	tfidf	10/10
rf	max * cluster	tfidf	10/10
rf	max * uncertainty	tfidf	10/10
rf	max * random	tfidf	10/10
rf	cluster * uncertainty	tfidf	10/10
rf	cluster * random	tfidf	10/10
rf	cluster	tfidf	5/5
rf	max	tfidf	5/5
rf	max * cluster	tfidf	5/5
rf	max * uncertainty	tfidf	5/5
rf	max * random	tfidf	5/5
rf	cluster * uncertainty	tfidf	5/5
rf	cluster * random	tfidf	5/5
rf	cluster	tfidf	5/10
rf	max	tfidf	5/10
rf	max * cluster	tfidf	5/10
rf	max * uncertainty	tfidf	5/10
rf	max * random	tfidf	5/10
rf	cluster * uncertainty	tfidf	5/10
rf	cluster * random	tfidf	5/10
rf	cluster	sbert	10/10
rf	max	sbert	10/10
rf	max * cluster	sbert	10/10
rf	max * uncertainty	sbert	10/10
rf	max * random	sbert	10/10
rf	cluster * uncertainty	sbert	10/10
rf	cluster * random	sbert	10/10
rf	cluster	sbert	5/5
rf	max	sbert	5/5
rf	max * cluster	sbert	5/5
rf	max * uncertainty	sbert	5/5
rf	max * random	sbert	5/5
rf	cluster * uncertainty	sbert	5/5

(continued)

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
rf	cluster * random	sbert	5/5
rf	cluster	sbert	5/10
rf	max	sbert	5/10
rf	max * cluster	sbert	5/10
rf	max * uncertainty	sbert	5/10
rf	max * random	sbert	5/10
rf	cluster * uncertainty	sbert	5/10
rf	cluster * random	sbert	5/10
rf	cluster	embeddingIdf	10/10
rf	max	embeddingIdf	10/10
rf	max * cluster	embeddingIdf	10/10
rf	max * uncertainty	embeddingIdf	10/10
rf	max * random	embeddingIdf	10/10
rf	cluster * uncertainty	embeddingIdf	10/10
rf	cluster * random	embeddingIdf	10/10
rf	cluster	embeddingIdf	5/5
rf	max	embeddingIdf	5/5
rf	max * cluster	embeddingIdf	5/5
rf	max * uncertainty	embeddingIdf	5/5
rf	max * random	embeddingIdf	5/5
rf	cluster * uncertainty	embeddingIdf	5/5
rf	cluster * random	embeddingIdf	5/5
rf	cluster	embeddingIdf	5/10
rf	max	embeddingIdf	5/10
rf	max * cluster	embeddingIdf	5/10
rf	max * uncertainty	embeddingIdf	5/10
rf	max * random	embeddingIdf	5/10
rf	cluster * uncertainty	embeddingIdf	5/10
rf	cluster * random	embeddingIdf	5/10
svm	cluster	doc2vec	10/10
svm	max	doc2vec	10/10
svm	max * cluster	doc2vec	10/10
svm	max * uncertainty	doc2vec	10/10
svm	max * random	doc2vec	10/10
svm	cluster * uncertainty	doc2vec	10/10
svm	cluster * random	doc2vec	10/10
svm	cluster	doc2vec	5/5
svm	max	doc2vec	5/5
svm	max * cluster	doc2vec	5/5
svm	max * uncertainty	doc2vec	5/5
svm	max * random	doc2vec	5/5
svm	cluster * uncertainty	doc2vec	5/5
svm	cluster * random	doc2vec	5/5
svm	cluster	doc2vec	5/10
svm	max	doc2vec	5/10
svm	max * cluster	doc2vec	5/10
svm	max * uncertainty	doc2vec	5/10

(continued)

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
svm	max * random	doc2vec	5/10
svm	cluster * uncertainty	doc2vec	5/10
svm	cluster * random	doc2vec	5/10
svm	cluster	tfidf	10/10
svm	max	tfidf	10/10
svm	max * cluster	tfidf	10/10
svm	max * uncertainty	tfidf	10/10
svm	max * random	tfidf	10/10
svm	cluster * uncertainty	tfidf	10/10
svm	cluster * random	tfidf	10/10
svm	cluster	tfidf	5/5
svm	max	tfidf	5/5
svm	max * cluster	tfidf	5/5
svm	max * uncertainty	tfidf	5/5
svm	max * random	tfidf	5/5
svm	cluster * uncertainty	tfidf	5/5
svm	cluster * random	tfidf	5/5
svm	cluster	tfidf	5/10
svm	max	tfidf	5/10
svm	max * cluster	tfidf	5/10
svm	max * uncertainty	tfidf	5/10
svm	max * random	tfidf	5/10
svm	cluster * uncertainty	tfidf	5/10
svm	cluster * random	tfidf	5/10
svm	cluster	sbert	10/10
svm	max	sbert	10/10
svm	max * cluster	sbert	10/10
svm	max * uncertainty	sbert	10/10
svm	max * random	sbert	10/10
svm	cluster * uncertainty	sbert	10/10
svm	cluster * random	sbert	10/10
svm	cluster	sbert	5/5
svm	max	sbert	5/5
svm	max * cluster	sbert	5/5
svm	max * uncertainty	sbert	5/5
svm	max * random	sbert	5/5
svm	cluster * uncertainty	sbert	5/5
svm	cluster * random	sbert	5/5
svm	cluster	sbert	5/10
svm	max	sbert	5/10
svm	max * cluster	sbert	5/10
svm	max * uncertainty	sbert	5/10
svm	max * random	sbert	5/10
svm	cluster * uncertainty	sbert	5/10
svm	cluster * random	sbert	5/10
svm	cluster	embeddingIdf	10/10
svm	max	embeddingIdf	10/10

(continued)

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
svm	max * cluster	embeddingIdf	10/10
svm	max * uncertainty	embeddingIdf	10/10
svm	max * random	embeddingIdf	10/10
svm	cluster * uncertainty	embeddingIdf	10/10
svm	cluster * random	embeddingIdf	10/10
svm	cluster	embeddingIdf	5/5
svm	max	embeddingIdf	5/5
svm	max * cluster	embeddingIdf	5/5
svm	max * uncertainty	embeddingIdf	5/5
svm	max * random	embeddingIdf	5/5
svm	cluster * uncertainty	embeddingIdf	5/5
svm	cluster * random	embeddingIdf	5/5
svm	cluster	embeddingIdf	5/10
svm	max	embeddingIdf	5/10
svm	max * cluster	embeddingIdf	5/10
svm	max * uncertainty	embeddingIdf	5/10
svm	max * random	embeddingIdf	5/10
svm	cluster * uncertainty	embeddingIdf	5/10
svm	cluster * random	embeddingIdf	5/10
lr	cluster	doc2vec	10/10
lr	max	doc2vec	10/10
lr	max * cluster	doc2vec	10/10
lr	max * uncertainty	doc2vec	10/10
lr	max * random	doc2vec	10/10
lr	cluster * uncertainty	doc2vec	10/10
lr	cluster * random	doc2vec	10/10
lr	cluster	doc2vec	5/5
lr	max	doc2vec	5/5
lr	max * cluster	doc2vec	5/5
lr	max * uncertainty	doc2vec	5/5
lr	max * random	doc2vec	5/5
lr	cluster * uncertainty	doc2vec	5/5
lr	cluster * random	doc2vec	5/5
lr	cluster	doc2vec	5/10
lr	max	doc2vec	5/10
lr	max * cluster	doc2vec	5/10
lr	max * uncertainty	doc2vec	5/10
lr	max * random	doc2vec	5/10
lr	cluster * uncertainty	doc2vec	5/10
lr	cluster * random	doc2vec	5/10
lr	cluster	tfidf	10/10
lr	max	tfidf	10/10
lr	max * cluster	tfidf	10/10
lr	max * uncertainty	tfidf	10/10
lr	max * random	tfidf	10/10
lr	cluster * uncertainty	tfidf	10/10

(continued)

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
lr	cluster * random	tfidf	10/10
lr	cluster	tfidf	5/5
lr	max	tfidf	5/5
lr	max * cluster	tfidf	5/5
lr	max * uncertainty	tfidf	5/5
lr	max * random	tfidf	5/5
lr	cluster * uncertainty	tfidf	5/5
lr	cluster * random	tfidf	5/5
lr	cluster	tfidf	5/10
lr	max	tfidf	5/10
lr	max * cluster	tfidf	5/10
lr	max * uncertainty	tfidf	5/10
lr	max * random	tfidf	5/10
lr	cluster * uncertainty	tfidf	5/10
lr	cluster * random	tfidf	5/10
lr	cluster	sbert	10/10
lr	max	sbert	10/10
lr	max * cluster	sbert	10/10
lr	max * uncertainty	sbert	10/10
lr	max * random	sbert	10/10
lr	cluster * uncertainty	sbert	10/10
lr	cluster * random	sbert	10/10
lr	cluster	sbert	5/5
lr	max	sbert	5/5
lr	max * cluster	sbert	5/5
lr	max * uncertainty	sbert	5/5
lr	max * random	sbert	5/5
lr	cluster * uncertainty	sbert	5/5
lr	cluster * random	sbert	5/5
lr	cluster	sbert	5/10
lr	max	sbert	5/10
lr	max * cluster	sbert	5/10
lr	max * uncertainty	sbert	5/10
lr	max * random	sbert	5/10
lr	cluster * uncertainty	sbert	5/10
lr	cluster * random	sbert	5/10
lr	cluster	embeddingIdf	10/10
lr	max	embeddingIdf	10/10
lr	max * cluster	embeddingIdf	10/10
lr	max * uncertainty	embeddingIdf	10/10
lr	max * random	embeddingIdf	10/10
lr	cluster * uncertainty	embeddingIdf	10/10
lr	cluster * random	embeddingIdf	10/10
lr	cluster	embeddingIdf	5/5
lr	max	embeddingIdf	5/5
lr	max * cluster	embeddingIdf	5/5
lr	max * uncertainty	embeddingIdf	5/5

(continued)

Model	Query Strategy	Feature extraction strategy	Training data [included/excluded]
lr	max * random	embeddingIdf	5/5
lr	cluster * uncertainty	embeddingIdf	5/5
lr	cluster * random	embeddingIdf	5/5
lr	cluster	embeddingIdf	5/10
lr	max	embeddingIdf	5/10
lr	max * cluster	embeddingIdf	5/10
lr	max * uncertainty	embeddingIdf	5/10
lr	max * random	embeddingIdf	5/10
lr	cluster * uncertainty	embeddingIdf	5/10
lr	cluster * random	embeddingIdf	5/10

Danka, Tivadar, and Peter Horvath. n.d. "modAL: A Modular Active Learning Framework for Python."

Le, Quoc V., and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." *arXiv:1405.4053 [Cs]*, May. http://arxiv.org/abs/1405.4053.

Lewis, David D., and Jason Catlett. 1994. "Heterogeneous Uncertainty Sampling for Supervised Learning." In *Machine Learning Proceedings 1994*, edited by William W. Cohen and Haym Hirsh, 148–56. San Francisco (CA): Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50026-X.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Zhang, Harry. 2004. "The Optimality of Naive Bayes." In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*. Vol. 2.