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.

Machine learning algorithms cannot predict the relevance of abstracts from the raw texts as they are. The content of the texts needs to be transformed into numerical representations. The processs of transforming texts to numerical feature vectors is called word embeddings.

A classical example of word embeddings is 'bag of words'. For each 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)

Word embeddings allows ASReview to predict relevance of abstracts from the features of abstracts of which relevance is known.

corpus = all the text:

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

The model is typically a learning algorithm used to predict the relevance of text.

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

Feature Extraction Strategies

split_ta = overall hyperparameter

TF-IDF The bag-of-words method is simplistic and will highly value often occuring but otherwise meaningless words such as “and”.

Term-frequency Inverse Document Frequency (**???**) circumvents this problem by adjusting a term frequency in a text with the inverse docuement frequency, the frequency of a given word in the entire corpus.

hyperparameters

```
ngram_max: int
        Can use up to ngrams up to ngram_max. For example in the case of
        ngram_max=2, monograms and bigrams could be used.
```

Doc2Vec Predicts words from context. Aims at capturing the relations between word (man-woman, king-queen). (Le and Mikolov 2014). Using a neural network.

using Continuous Bag-of-Words (CBOW), Skip-Gram model, …. Word vector W and extra: document vector D, trained to predict words in the text.

From gensim (**???**).

```
Arguments
---------
vector_size: int
    Output size of the vector.
epochs: int
    Number of epochs to train the doc2vec model.
min_count: int
    Minimum number of occurences for a word in the corpus for it to
    be included in the model.
workers: int
    Number of threads to train the model with.
window: int
    Maximum distance over which word vectors influence each other.
dm_concat: int
    Whether to concatenate word vectors or not.
    See paper for more detail.
dm: int
    Model to use.
    0: Use distribute bag of words (DBOW).
    1: Use distributed memory (DM).
    2: Use both of the above with half the vector size and concatenate
    them.
dbow_words: int
    Whether to train the word vectors using the skipgram metho
```

SBERT BERT-base model with mean-tokens pooling (**???**)

embeddingIdf This model averages the weighted word vectors of all the words in the text, in order to get a single feature vector for each text. The weights are provided by the inverse document frequencies

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

A number of decision trees are fit on bootstrapped samples of the original data, (**???**) RandomForestClassifier from sklearn

Arguments ——— n_estimators: int Number of estimators. max_features: int Number of features in the model. class_weight: float Class weight of the inclusions. random_state: int, RandomState Set the random state of the RNG. """

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

amount of training data

- n_instances = number of papers queried each query
- n_queries = number of queries
- n_prior_included: 5
- n_prior_excluded:

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.

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

References

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.