Classification of documents using dataless technique

Muqeeth

Indian Institute of Technology Madras

Abstract. The labels of the documents carry a meaning.We human beings categorize documents based on meaning of the labels. Such technique is called dataless classificaion. In this project we present an approach of classifying documents among different labels using the meaning of the labels.This approach does not involve training as in the case of supervised learning. The key idea is to build a semantic interpreter. In the semantic space constructed, we represent the document to classify and all the labels. The label which is closer to the document is assigned to that document.

Keywords: Explicit semantic analysis · labels · 20newsgroup dataset.

1 Introduction

1.1 Importance

Classification of files is important to manage them. For example, files in our system are categorized into folders like downloads,desktop,documents etc which helps in easier retrieval.The emails we receive can be classified as academic,sport etc .Supervised techniques extract features from train documents and use them to classify test documents. If the test document is of different domain from train documents, supervised techniques face issues with it.The dataless technique will help to resolve such problem if we have rich semantic space.When sufficient training examples are not present, the dataless technique will classify documents using label meanings.Then human judgement can be made on those documents and sufficient training examples can be created which are further used for supervised learning.

The performance of dataless technique depends on semantic space constructed.It is shown that explicit semantic analysis on wikipedia constructs a rich semantic space.Since wikipedia has different domains such as politics,logic,culture etc. Classification of documents in all domains would be easier if we use wikipedia.The description of labels will also play an important role in classificaion.For example, we want to classify documents among NLP and CV, label description such as artificial intelligence,machine learning respectively would not be helpful,but labels as text processing,image processing would be helpful to classify them.

1.2 20newsgroup

20newsgroup dataset has about 20,000 emails categorized under 20 labels with each label having about 1000 emails.There are six top level labels in this dataset.Since these are emails,text processing such as removing headers(FROM: SUBJECT:),footer(WITH REGARDS:),quotes are done.The documents are then processed to form bag of words. The documents are then shuffled and 8000 documents with each label having about 400 documents is chosen as test set for implementing dataless classification technique. Fig. 1).

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Fig. 1. Categories in 20newsgroup dataset.

2 Methodology

2.1 Semantic Interpreter

Simple english wikipedia is chosen to construct semantic space.The xml dump is parsed to obtain 0.2 million articles of simple wikipedia.while parsing , articles with less than 100 words and less than 5 outgoing and incoming links are removed.The stop words like the,for,from etc which are used frequently in the articles are removed.The resulting articles are then lemmatized using wordnet lemmatizer and tokenized using porter stemmer to obtain 0.5 million tokens.Normalized tfidf scheme is used to term-document matrix.In normalized tfidf the termfrequency is count of term in the document divided by maximum count of word in any given document.The constructed semantic space (space spanned by articles of wikipedia) is used to represent document to be classified and all labels.

2.2 Implementation

Each word is represented as a vector of size given by number of articles in wikipedia. The vector representation of the document is given by weighted sum

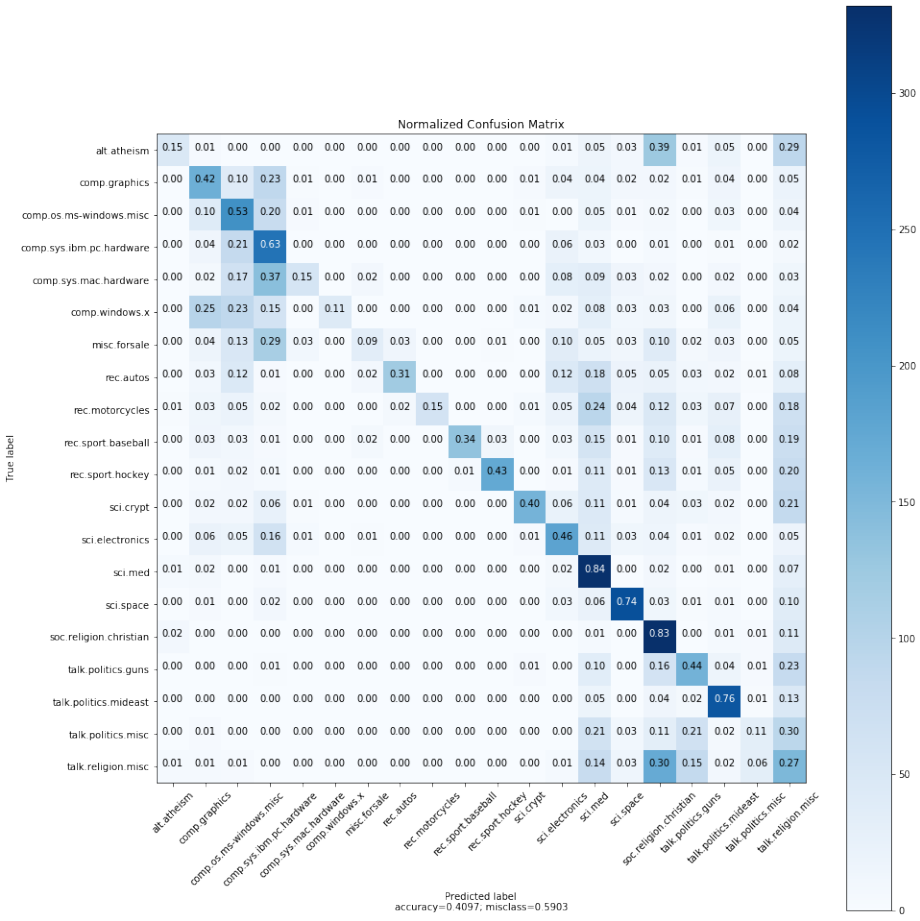

Fig. 2. Normalized confusion matrix

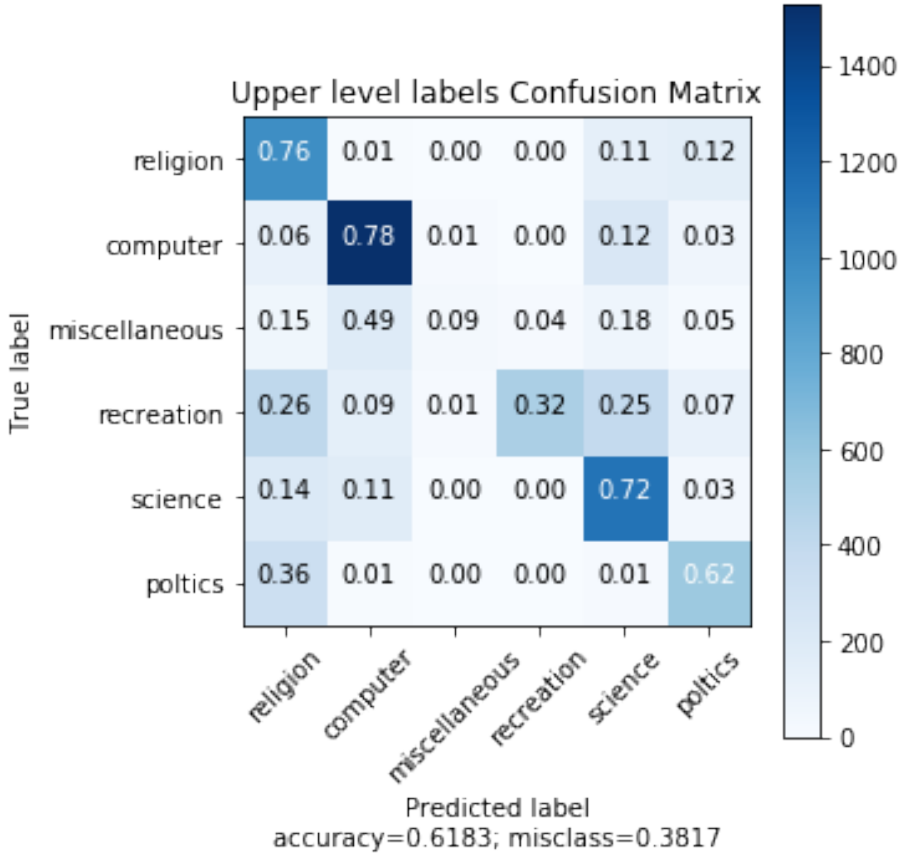

Fig. 3. six top level labels confusion matrix

of vectors of the words in the document.The weights are given by tf-idf.tf is term frequency of word in that document and idf is inverse document frequency of word in wikipedia articles.The labels are given the short description ex: sci.space as science space earth . which help in finding their vector representation similar to what we have done for documents.Table 1,the description of the labels are given. The vectors for labels should be as orthogonal as possible for good classi-

Table 1. Description of labels in 20newsgroup dataset.

Label	Description
alt.atheism	atheist atheism god islamic
comp.graphics	format animation polygon graphics
comp.os.ms-windows.misc	windows dos microsoft ms driver card printer
comp.sys.ibm.pc.hardware	bus pc motherboard bios board computer dos
comp.sys.mac.hardware	mac apple powerbook
comp.windows.x	window motif xterm sun windows
misc.forsale	sale discount price
rec.autos	car ford auto toyota honda nissan bmw
rec.motorcycles	bike motorcycle yamaha
rec.sport.baseball	baseball ball hitter
rec.sport.hockey	hockey wings espn
sci.crypt	encryption key crypto algorithm security
sci.electronics	circuit electronics radio signal battery
sci.med	medical disease patient
sci.space	earth solar spacecraft lunar shuttle moon launch nasa orbit space
soc.religion.christian	christian god christ church bible jesus
talk.politics.guns	gun fbi guns weapon compound
talk.politics.mideast	israel arab jews jewish muslim
talk.politics.misc	homosexuals gay libertarians tax
talk.religion.misc	christian morality jesus god religion horus

fication. In Table 2,the order of similarities between close labels is first decimal and between not close labels is second and third decimal.The cosine similarity between vector of document and labels is used for classification.The label with maximum cosine similarity is assigned to the given document.

Table 2. similarity between labels in 20newsgroup dataset.

	soc.religion.christian	sci.space	rec.sport.hockey	comp.windows.x
alt.atheism	0.26	0.020	0.0021	0.018
sci.electronics	0.008	0.063	0.025	0.016
rec.sport.baseball	0.004	0.012	0.023	0.008
comp.os.ms-windows.misc	0.024	0.028	0.009	0.18

2.3 Results and analysis

The metrics used to evaluate our model are accuracy,macrof1 score and microf1 score.The accuracy is ratio of number of documents correctly labelled to total number of documents. The precision and recall for all 20 labels are calculated.microf1 score is f1 score of average precision and average recall.macrof1 score is average of f1 score of each label in dataset.The results are presented in Table 3.The precision and recall for all 20 labels in two techniques are presented in Table 4 Binary classification is done on chosen set of labels and results are

Table 3. Metrics

	Supervised SGD classfier	Dataless technique
Accuracy	0.695	0.409
macro f1	0.671	0.481
microf1	0.684	0.394

Table 4. Recall and precision on all labels

Label	Precision DC	Recall DC	Precision SGD	Recall SGD
alt.atheism	0.706	0.150	0.610	0.388
comp.graphics	0.383	0.421	0.702	0.668
comp.os.ms-windows.misc	0.318	0.532	0.648	0.642
comp.sys.ibm.pc.hardware	0.290	0.627	0.711	0.647
comp.sys.mac.hardware	0.703	0.148	0.751	0.714
comp.windows.x	0.954	0.106	0.773	0.762
misc.forsale	0.515	0.088	0.739	0.815
rec.autos	0.846	0.305	0.796	0.729
rec.motorcycles	0.983	0.150	0.521	0.791
rec.sport.baseball	0.957	0.337	0.817	0.801
rec.sport.hockey	0.920	0.434	0.807	0.924
sci.crypt	0.872	0.396	0.722	0.757
sci.electronics	0.427	0.460	0.652	0.483
sci.med	0.313	0.837	0.752	0.797
sci.space	0.673	0.740	0.704	0.786
soc.religion.christian	0.324	0.834	0.588	0.851
talk.politics.guns	0.424	0.436	0.585	0.678
talk.politics.mideast	0.549	0.755	0.766	0.827
talk.politics.misc	0.434	0.106	0.672	0.390
talk.religion.misc	0.147	0.27	0.46	0.119
Averages	0.587	0.407	0.689	0.678

presented in Table 5 and Table 6 for both supervised and dataless techniques. Normalised confusion matrix for all labels in dataset is shown in figure 2.We

Table 5. Accuracy for binary classification between labels

	soc.religion.christian	rec.sport.hockey	sci.space	comp.windows.x
alt.atheism	0.622	0.791	0.858	0.934
sci.electronics	0.886	0.837	0.894	0.711
rec.sport.baseball	0.702	0.723	0.748	0.874
comp.os.ms-windows.misc	0.948	0.937	0.946	0.572

Table 6. Accuracy for binary classification between labels using supervised technique

	soc.religion.christian	rec.sport.hockey	sci.space	comp.windows.x
alt.atheism	0.811	0.924	0.869	0.949
sci.electronics	0.945	0.960	0.885	0.918
rec.sport.baseball	0.942	0.914	0.919	0.957
comp.os.ms-windows.misc	0.943	0.953	0.909	0.860

achieved an accuracy of 41 percent.20 labels in dataset are made into six labels and the confusion matrix for it is shown in figure 3.We achieved an accuracy of 62 percent.

3 Discussion and conclusion

For the binary classification among chosen set of labels both the supervised and dataless techniques are performing better.Dataless technique has low accuracy as it performed badly on computer and recreation classes.In computer category,the model is getting confused among its sub categories.We can see cluster formed for computer category in figure 2.The label description for computer class sub categories might be a reason for this.Recreation category words are not having a good representation using simple wiki interpreter. Dataless technique has low average recall which shows a need for rich sematic interpreter.The average precision for both the techniques are close by which shows we have good description of labels. Accuracy could be improved if we would have used full wikipedia for constructing semantic space.

References

1. Author, Ming-Wei Chang, Lev Ratinov, Dan Roth and Vivek Srikumar:Importance of Semantic Representation: Dataless Classification. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)
2. Author,Yangqiu Song and Dan Roth:On Dataless Hierarchical Text Classification. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence
3. Home page for 20newsgroups dataset, http://qwone.com/ jason/20Newsgroups/.