Document classification using 3-view of document representations and ensemble : TF-IDF, LDA and Doc2Vec

Eunji Jun (2016021311)
Deokseong Seo (2016020548)
Honggyu Jung (2016010933)

{*ejjun*92, *heyhi*16, *hkjung*00}@korea.ac.kr

Department of Brain and Cognitive Engineering,
Korea University

14th June 2017

Contents

Introduction

Background

- Use document classification algorithms to categorize data to quickly and efficiently locate documents and reduce storage and backup costs

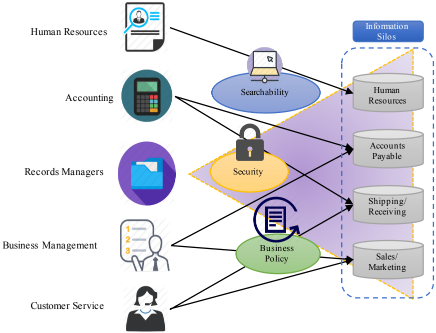

Figure 1: Document classification

Motivation

- As the amount and size of data increases, the necessity of classification and organization of documents increases.
  - Effective classification strategy is needed.
- Multi-view learning method does not exist.

Proposed method

Proposed method

- Propose an ensemble model that combines all three document representation methods (TF-IDF, LDA and Doc2Vec)
  - Compare the performance using the proposed ensemble model with the individual models, respectively
  - Compare different classifiers used for document classification.

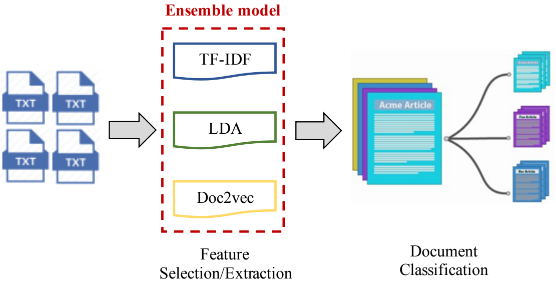

Figure 2: Illustration of the proposed method

Document representation

- TF-IDF(Term Frequency-Inverse Document Frequency)
  - Generate the weight of each word based on the appearance frequency and uniqueness
  - Extract an 100 dimensional vector for each document

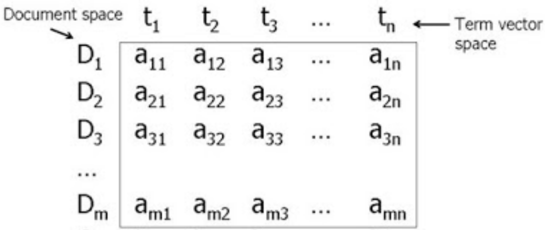

Figure 3: TF-IDF

- LDA(Latent Dirichlet allocation)
  - Estimate the distribution of topics in a document and the distribution of words
  - Extract an 100 dimensioanl vector for each document

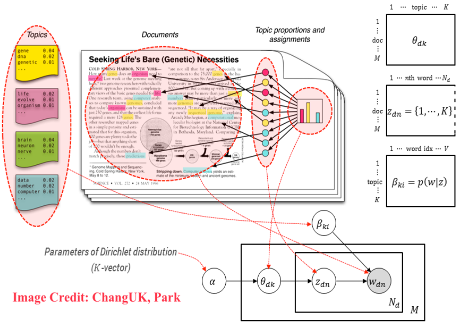

Figure 4: LDA

- Doc2Vec(Document to Vector)
  - Words and a document ID are used to extract an 100 dimensional vector through the backpropagation learning method of neural network.

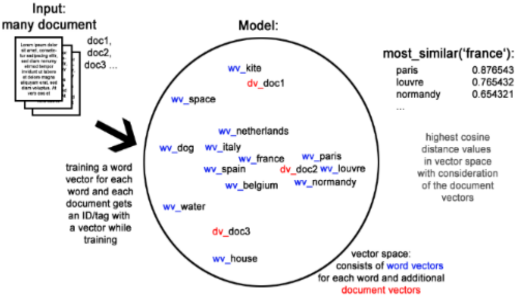

Figure 5: Doc2Vec

Document classification

- Naive Bayes classifier
  - A probabilistic classifier which computes the probability of a document d being in a class c
- Decision tree
  - A tree in which the internal nodes are labeled by the features, the edges leaving a node are labeled by tests on the feature's weight, and the leaves are labeled by categories

Experiment

Dataset

- Document dataset in different fields

Table 1: Data description

Dataset	Description	Range	Row	Source
Economic	Whether a news article data is associated with the US economy	No : 6,458 (82.12%) Yes : 1,406 (17.88%)	7,864	http://www.crowdflower.com/data-for-everyone
Ohsumed	Articles related abstracts of medical data	C04 : 2,630 (50.77%) C14 : 2,550 (49.23%)	5,180	http://disi.unitn.it/moschitti/corpora.htm
Reuters	Documents obtained by the Reuters news data	Earn : 3,953 (51.67%) Non-earn : 4,697 (48.33%)	7,650	http://www.daviddlewis.com/resources/testcollections/reuters21578/

Experiment procedure

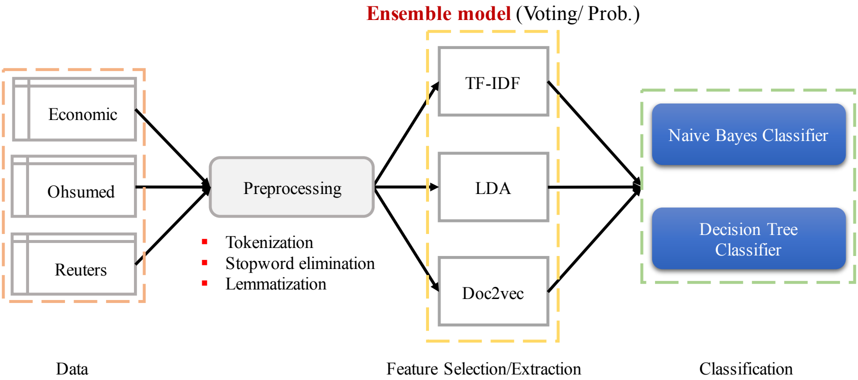

Figure 6: Experiment procedure

Results

Results

- For Economic dataset,

Table 2: Data description

Classifier	Representation	Accuracy(%)	Recall(%)	Precision(%)	F1-measure(%)
Naive Bayesian	TF-IDF	63.68±0.87	68.07±1.97	28.57±1.09	40.23±1.19
	LDA	50.85±1.08	73.80±2.07	22.93±1.06	34.98±1.33
	Doc2Vec	75.56±0.93	40.79±1.86	34.71±1.83	37.47±1.46
	Ensemble(Voting)	79.64±8.83	4.02±13.33	18.35±16.02	3.24±6.04
	Ensemble(Prob.)	78.35±10.06	6.24±16.15	18.37±17.60	4.40±8.04
Decision tree	TF-IDF	73.45±0.86	27.37±2.38	26.53±1.94	26.91±1.87
	LDA	70.95±0.91	23.21±2.19	21.28±1.74	22.17±1.76
	Doc2Vec	71.00±0.86	23.17±2.04	21.53±1.77	22.31±1.75
	Ensemble(Voting)	80.14±7.06	3.10±9.95	17.14±12.85	3.02±5.09
	Ensemble(Prob.)	78.97±10.01	5.39±16.12	22.26±18.91	3.85±7.00

- For Ohsumed dataset,

Table 3: Data description

Classifier	Representation	Accuracy(%)	Recall(%)	Precision(%)	F1-measure(%)
Naive Bayesian	TF-IDF	86.85±0.71	85.89±1.05	87.17±1.17	86.52±7.79
	LDA	75.20±0.91	77.04±1.50	73.70±1.37	75.32±1.02
	Doc2Vec	65.31±1.26	59.64±1.89	66.58±1.94	62.90±1.60
	Ensemble(Voting)	52.92±3.25	31.42±30.84	57.53±11.29	31.45±22.48
	Ensemble(Prob.)	52.66±2.35	27.00±29.03	57.64±11.93	28.35±21.16
Decision tree	TF-IDF	86.66±0.77	85.85±1.23	86.87±1.07	86.35±0.80
	LDA	75.16±0.95	77.23±1.44	73.60±1.43	75.36±1.01
	Doc2Vec	65.54±1.35	59.72±1.84	66.94±1.89	63.10±1.51
	Ensemble(Voting)	52.66±2.83	33.56±33.09	58.80±11.59	32.47±21.72
	Ensemble(Prob.)	52.45±2.61	29.73±31.33	56.94±9.94	29.93±21.50

- For Reuters dataset,

Table 4: Data description

Classifier	Representation	Accuracy(%)	Recall(%)	Precision(%)	F1-measure(%)
Naive Bayesian	TF-IDF	94.24±0.39	97.25±0.53	92.69±0.73	94.91±0.35
	LDA	82.48±0.67	79.88±1.14	87.38±0.99	83.45±0.72
	Doc2Vec	65.72±0.76	56.69±1.18	75.14±1.18	64.61±0.92
	Ensemble(Voting)	53.37±5.41	29.85±18.96	71.14±10.74	38.45±14.83
	Ensemble(Prob.)	53.26±4.80	28.89±18.18	72.18±10.26	37.76±14.57
Decision tree	TF-IDF	94.19±0.40	97.28±0.54	92.57±0.74	94.86±0.36
	LDA	82.61±0.70	79.97±1.11	87.48±0.95	83.55±0.74
	Doc2Vec	65.70±0.84	56.78±1.27	75.15±1.31	64.68±1.03
	Ensemble(Voting)	54.23±5.42	30.82±19.28	73.93±9.82	39.76±14.69
	Ensemble(Prob.)	53.83±5.15	29.96±19.43	72.43±10.66	38.66±15.09

Conclusion

The performance of the two classifiers is similar, except for the economic dataset.

Effective document representation differs for each dataset.

Most individual models outperform ensemble models, but only accuracy in the economic dataset.

References

[1] Zhou,Z-H., & Li, M. (2005) “Tri-Training: Exploiting Unlabeled Data Using Three Classifiers”, IEEE Transactions On Knowledge And Data Engineering, 17(11),p1529-1541

[2] Ranzato, M., & Szummer, M. (2008) “Semi-supervised leaning of compact document representations with deep networks”, In Proceedings of the 25th international conference on Machine learning, p792—799

[3] Wang, D., Thint, M., & Al-Rubaie, Ahmad. (2012) “Semi-Supervised Latent Dirichlet Allocation and Its Application for Document Classification”, In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligent Agent Technology Volume 03, p 306-310

[4] Lu, Y., Okada, S., & Nitta, K., (2013) “Semi-supervised Latent Dirichlet Allocation for Multi-label Text Classification” , 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2013 Amsterdam, The Netherlands, p 351-360

[5] Zhang Y., & Wei, W. (2014) “A jointly distributed semi-supervised topic model”, Neurocomputing, Volume 134, p38-45

[6] Le, Q.V., Milolov, T., (2014) Distributed Representations of Sentences and Documents “, In Proceedings of the 31st International Conference on Machine Learning, Beijing, China,2014. JMLR: W&CP Volume 32,

[7] Wu, X., Fang L., Wang, P., & Yu, N.(2015) “Performance of Using LDA for Chinese News Text Classification”, In Proceeding of the IEEE 28th Canadian Conference on Electrical Computer Engineering Halifax, Canada, p1260-1264

Thank you
for your attention

(Q & A)