Project 5 – Advanced Data Mining Applications

CS548 / BCB503 Knowledge Discovery and Data Mining - Fall 2017

Prof. Carolina Ruiz

Student: **Yuchen Shen**

Description of the particular problem within the selected data mining topic to be addressed in this project	/15
Description of the approach used in this project to tackle the above problem. *All data mining techniques you use in this project for pre-processing, mining and evaluation must have been covered in class during this semester.*	/25
Description of the dataset selected	/15
Appropriateness of the dataset selected with respect to this topic/problem	/10
Guiding questions	/10
Preprocessing	/10
Experiments: • Sufficient & coherent	/25
• Objectives, Data, Additional Pre/Post-processing	/20
• Presentation of results	/20
• Analysis of results	/30
Overall discussion, comparisons, and conclusions	/20
TOTAL	/200

Total Written Report: _____/200 = _____/100

Class Presentation: _____/100

Class participation during project presentation: _____/100

Do not exceed the given page limits for this written report

Topic: Text Mining---News Domain Detection Using Classification and Clustering <at most 1 page>

1. **Description of the particular problem within the selected data mining topic to be addressed in this project:**

In order to quickly post news on the right part of web or newspaper, knowing the domain of a particular news is important; Knowing what news domain are most similar to each other; Knowing what domain of news is most difficult to recognize.

2. **Description of the approach used in this project to tackle the above problem:**

Using Text Mining skills to change text data in News into computer readable data; Using Classification methods to find patterns of different type of News including World, Sports, Business and Sci/Tech and detect new coming News type; Trying Clustering methods to find similar domain and news which can be clustered in many different fields.

3. **Dataset Name: Ag's News**
4. **Where found: Deep Learning Open Dataset https://deeplearning4j.org/cn/opendata**
5. **Dataset Description:**

The data is from AG's News Dataset, which is constructed from a collection of 1 million news articles. I use the 'test data set' which have a suitable number of samples. The number of news samples in this used dataset is 7,600, for each news type is World, Sports, Business and Sci/Tech. There is no missing data in this set.

The csv file contains all the news samples with 2 columns corresponding to news type and news content. The contents are escaped using double quotes ("), and any internal double quote is escaped by 2 double quotes (""). New lines are escaped by a backslash followed with an "n" character, that is "\n". The probabilities of each type of news are same in the whole dataset, they are uniformly distributed, and enhanced the precision of the analysis.

6. **Initial data preprocessing, if any:**

Split the whole dataset into training data and testing data. Extract 475 sample from each type of news from the 'train dataset' to construct the new test dataset which contains 1900 news samples. Delete the attribute 'title' which represent a short description of the news and I will not use it. As The probabilities of each type of news are same in the whole dataset, some classification method will be excluded, such as ZeroR.

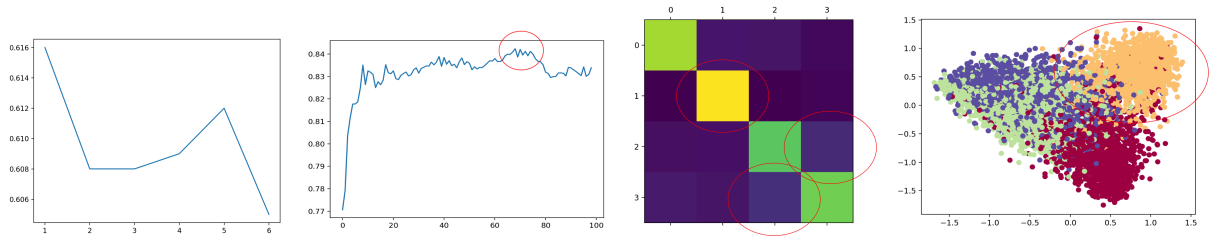
'**Three Guiding Questions about the dataset domain:**

1. Which news domain will always use domain words?
2. Which domains using words are most similar with each other?
3. Which news domain uses widest word domain?

Summary of Experiments. *At most 2 page.*							
Tool	**Pre-process**	**Mining Technique**	**parameter**	**result (show particular wrong classify result)**	**Time taken**	**Evaluation**	**Observations about experiment Observations about visualization Interpretation results**
python	Tokenizing Remove stopword; Phrases: (1,1)	Tf-idf text mining;	Min appear: 1% Max appear: 90%	384 features	0.0012s	In decision tree, score 0.592	Change the range of feature appear rate, a lower boundary can include domain words for each domain. Use (0.001~0.8) as appear range.
python			Min appear: 0.1% Max appear: 80%	3724 features	0.001s	In decision tree, score 0.616	
Python	Tokenizing Remove stopword; Tf-idf Transform; Min appear 0.1%; Max appear 80%;	Text Mining with Tfidf; OneR	_____	Only 'Sports' most correctly classified.	Tf-idf:1.20s OneR:0.00s	Precision:0.26 F-score:0.25 Recall:0.23	OneR classifier give a bad result as ZeroR, because a single word will not represent a news.
Python		Text Mining with Tfidf; Decision tree;	Phrase: (1,2) min_sample: 0.01 Euclidean	'Business' -> 'Sci/Tech':32%	Tf-idf:1.90s Tree:0.00s	Precision:0.61 F-score:0.61 Recall:0.61 Score: 0.608	Only contain single word will make classifier a bit better, so choose single word; 'Business' type has a high wrong classify rate and always classified to 'Sci/Tech'.
python			Phrase: (1,1) min_sample: 0.01 Euclidean	'Business' -> 'Sci/Tech':25% 'Sci/Tech' -> 'Business':23%	Tf-idf:1.20s Tree:0.00s	Precision:0.62 F-score:0.62 Recall:0.61 Score: 0.616	
Python		Text Mining with Tfidf; Random Forest;	10 trees; max_depth:50 phrase:(1,1) Euclidean	'Business' -> 'Sci/Tech':23% 'World' ->'Sport':14%	Tf-idf:1.20s Tree:0.71s	Precision:0.69 F-score:0.66 Recall:0.64	Have a better performance than decision tree, but still have a high rate of wrong classified 'Business' and 'Sci/Tech'; large max_depth make a better result;
Python			10 trees; max_depth:10 phrase:(1,1) Euclidean		Tf-idf:1.20s Tree:0.02s	Precision:0.65 F-score:0.64 Recall:0.64	
python		Text Mining with Tfidf; K Nearest Neighbor;	kNeighbor:1 Phrase:(1,1) Euclidean	'Business' -> 'Sports':17%	Tf-idf:1.20s Knn:0.00s	Precision:0.77 F-score:0.77 Recall:0.77	'Sports' domain always has a good prediction, sports news always has domain word that can be easily
python			KNeighbor:62 Phrase:(1,1) Euclidean	'Business -> 'Sci/Tech':21%	Tf-idf:1.20s Knn:0.01s	Precision:0.83 F-score:0.83 Recall:0.83	

python			Kneighbor:100 Phrase:(1,1) Euclidean	'Sci/Tech' -> 'Business':19%	Tf-idf:1.20s Knn:0.01s	Precision:0.82 F-score:0.83 Recall:0.82	recognized; A suitable K Neighbor will improve classification result.
Weka	Tf-idf Transform; Tokenizing; Stemmer; Remove stopword;	Text Mining with Tfidf; decision tree(J48);	Leave 200 attributes; Min number per leaf: 10; Euclidean	No special wrong classify result, wrong rates are even.	Model/test 16.22s	Precision:0.57 RMSE:0.33	'Business' type has a high wrong classify rate and always classified to 'Sci/Tech'. guess words in these areas are always similar.
Weka	Tf-idf Transform; Tokenizing; Stemmer; Remove stopword;	Text Mining with Tfidf; Random Forest;	Leave 200 attributes; Feature: log_2(#predictors) + 1; Euclidean	No special wrong classify result, wrong rates are even.	Model/test 32.56s	Precision:0.59 RMSE:0.32	Little improve compare to J48, a bigger leaving attribute can have better result, but Weka unable do that.
python	Tokenizing Remove stopword; Tf-idf Transform; Min appear 0.1%; Max appear 80%;	Text Mining with Tfidf; K-means clustering;	K = 4 (4 initial) domain Euclidean	'World' and 'Sports' News have own cluster, 'Business' mix with others.	0.06s	Accuracy: 0.279 SSE: 152.94 Purity:0.336 NMI:0.241	k-means not works well in text data, but result shows 'Sport' news always cluster together.
python			K = 6 (find what will be cluster in 2 new clusters) Euclidean	'World' News classify separately	0.12s	SSE: 130.30 Purity:0.427 NMI:0.262	Add two more cluster than 4 domains, the result shows "World' news clustered in 3 clusters.
python		Tfidf; Hierarchical clustering	K = 4 (initial 4) Euclidean	Almost news stays in a single cluster.	0.781s	SSE:1328.62 Purity:0.173 NMI:0.154	After Tfidf, text data are much different with each other
python		Text Mining with Tfidf; DBSCAN clustering;	Eps = 1.0 Min point:100 Euclidean	Change eps from 0.1 to 2.0 find the best result	0.032s	Purity: 0.212	DBSCAN better than hierarchical but not suitable with text data.
python			Eps = 1.325 Min point:100 Euclidean		0.032s	Purity:0.247	
Weka	Tf-idf Transform; Tokenizing; Stemmer; Remove stopword;	Text Mining with Tfidf; K-means clustering;	Leave 200 attributes; K = 4;	Most samples clustered in 3 clusters and another cluster only contains 1 'world' news	0.66s	SSE: 33794.37 Incorrect: 70.9%	'World' has a wide words domain, these news always use word different from other 'World' news.

Analysis of Results: (at most 1 page) 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include visualizations.



1. Using text mining method, when we change parameter min and max words appear rate, a lower min rate will increase the amount of words which can be included; When I change the min from 1% to 0.1%, selected word vector from 384 to 3724; Lower boundary will include words which can represent each domain. As the first picture shows, Tf-idf method sorts words and phrases. The picture shows features include different phrases range, in these case, features just include single words(score:0.616) or include 5 words phrases(score:0.612) performance best. I use words appear range from 0.1% to 80% and just include single words.
2. For different classification method, the result shows similar pattern and all gives good performance. Among decision tree, random forest and K Nearest Neighbor methods, Knn gives the best result which have precision above 80%, and Knn can easily classify most news automatically. As the second picture above, when searching 62 nearest neighbor will make the Knn classification method performance best, but will not change too much. The yellow part of the third picture represent the 'Sports' news, which has the highest precision and means 'Sports' domain can be easily recognized rather than other domains. For guide question 1, as I think in advance, 'Sports' news will always use domain words in different sports area, because words like 'soccer' and 'ball' will not always appear in other domains. For guide question 2, at bottom right of the picture, 2 and 3 represent 'Business' and 'Sci/Tech' news, they both have a high rate of incorrectly classified into the other domain which means they use similar words and the difference between these 2 domains is not quite clear.
3. Clustering methods all gives bad results with accuracy below 30%, traditional clustering methods is suitable with text data since there will be too many attributes, but using manifold visualization method, we can see in the 4th picture which is the real domain distribution, the orange part represent 'Sports' news which cluster mostly together; 'Business'(green) and 'Sci/Tech'(blue) mixed together, these two domains do not have a clear difference; For guide question 3, 'World'(red) have its own cluster, but it has many point mixed with other domains. 'World' domain news use the widest words.

Summary of what you learned in this project:

In text mining, we need to have wide range of words appear rate and sometime phrases rather than just single word can make a text more recognizable; K Nearest Neighbor performs well in news data and a well chose K will improve the model; Traditional clustering methods not suit with text data.