http://dx.doi.org/10.14400/JDC.2014.12.6.325

Classification Accuracy by Deviation-based Classification Method with the Number of Training Documents

Yong-Bae Lee
Dept. of Computer Education, Jeonju National University of Education

학습문서의 개수에 따른 편차기반 분류방법의 분류 정확도

이용배
전주교육대학교 컴퓨터교육과

Abstract It is generally accepted that classification accuracy is affected by the number of learning documents, but there are few studies that show how this influences automatic text classification. This study is focused on evaluating the deviation-based classification model which is developed recently for genre-based classification and comparing it to other classification algorithms with the changing number of training documents. Experiment results show that the deviation-based classification model performs with a superior accuracy of 0.8 from categorizing 7 genres with only 21 training documents. This exceeds the accuracy of Bayesian and SVM. The Deviation-based classification model obtains strong feature selection capability even with small number of training documents because it learns subject information within genre while other methods use different learning process.

Key Words : automatic classification, accuracy of classification, the number of training documents

요 약 일반적으로 자동분류는 학습문서의 개수에 영향을 받는다고 알려져 있지만 실제로 학습문서의 수가 텍스트 자동분류에 어떻게 영향을 주는지 입증한 연구는 거의 없었다. 본 연구에서는 학습문서 수가 자동분류에 어떤 영향을 주는지 알아보기 위해 최근에 개발된 편차기반 분류방법을 중심으로 다른 분류 알고리즘과 비교하는데 초점을 두었다. 실험결과, 편차기반 분류모델은 학습문서의 수가 총 21개(7개 장르)인 상황에서 정확도가 0.8로 베이지안이나 지지벡터기계보다 우수하게 나타났다. 이것은 편차기반 분류모델이 장르내의 주제정보를 이용하여 학습하기 때문에 학습문서의 수가 적더라도 다른 학습방법보다 좋은 자질 선택 능력을 갖는다는 것을 입증한 것이다.

주제어 : 자동 분류, 분류 정확도, 학습문서의 개수

1. Introduction

1.1 Aim of the study

Text classification is one of the most important components in the field of machine learning and data analytics. The need for Big data processing has been increased and more and more automatic classification researchers are dealing with developing and evaluating

Received 13 April 2014, Revised 19 May 2014
Accepted 20 June 2014
Corresponding Author: Yong-Bae Lee
(Jeonju National University of Education)
Email: yblee@jnue.kr

ISSN: 1738-1916

new classification models.

Traditional text classification used to be applied in various areas. For example, distributing inquiries that are listed at a customer center to the person in charge, grouping retrieval results in subject categories and recognizing the language or the author based on the text styles. Recently it enables broader application such as filtering[1] web documents to find a target genre or analyzing[2] preference tendency based on on-line reviews on movies or restaurants to make a decision.

Classification accuracy is the most critical factor that is used to evaluate classification models. It is generally accepted that classification accuracy is affected by the quality and the number of learning documents, the number of categories and the type and the number of features. Interestingly, there are few studies that show how these variables influence automatic classification.

This study is performed through systematically designed experiments to analyze the influence of the number of training documents on classification accuracy with objectivity. Special emphasis is put on the different performance of the deviation-based classification model which has recently developed comparing to the accuracy of other classification algorithms with the change of the number of training documents.

1.2 Related works

Fuxman[3] pointed out the significant effect of training documents with high quality on automatic categorization. He suggested methods to select websites which are well known and contain massive data so that we do not need to collect training documents manually. He also suggested a method to input queries on these websites to extract training documents automatically.

Maeda[4] noted that the small number of training documents lowers classification accuracy. As a resolution to decrease classification errors that occur when there are fewer training documents, Maeda revised the calculation method of term weights in Bayesian algorithm. In Maeda's experiment this method showed about 0.5 of accuracy with 40 to 50 training documents.

Apte[5] introduced three important factors in text classification. They are the number of available training documents, the number of useful features in extracted words and the number of categories that were used for training. Apte classified Reuter's news data with an improved decision tree classification method and a Bayesian method. Decision tree classification method provided excellent results in this experiment.

Recent studies are corroborating the idea that the number of training documents affects the result of automatic classification. Still they are not enough to support the idea of the different performance of classification methods with various numbers of training documents.

The rest of this paper is organized as follows. Chapter 2 identifies the characteristics of classification models that are participating in evaluation. Chapter 3 analyses the changes in the classification accuracy that are gained from various numbers of training documents. The last chapter offers a conclusion and recommendations for future study.

2. Classification models for evaluation

2.1 Deviation-based classification method

Deviation-based classification method(DCM)[6,7] was developed to classify digital documents on a genre basis. Genre-based classification is not classifying the subject or the topic of documents; instead, it classifies based on the type or the style of documents. The basic concept that DCM supports is that there can be many subject classes within a genre and a good genre-revealing term would show up across different subject classes while appearing in many documents in the genre.

$$RVal_g(t_k) = DF_g(t_k) \bullet (1 - \sqrt{\frac{\sum_{i=1}^{n_s}(DF_g(t_k) - DF_g(t_k^i))^2}{n_s}}) \quad (1)$$

$$DVal_g(t_k) = \sqrt{\frac{\sum_{i=1}^{n_g}(RVal_g(t_k) - RVal_i(t_k))^2}{n_g}} \quad (2)$$

- $DF_g(t_k)$: term t_k's document frequency ratio for genre g
- $DF_g(t_k^i)$: term t_k's document frequency ratio for subject class i within the genre g
- n_s : the number of subject classes within the genre g
- n_g : the number of genre classes

$RVal_g(t_k)$ (1) shows the procedure to calculate the weighted term representing genre. It can be calculated by multiplying the subject class deviation of the term with its frequency ratio. Centroid vector $DVal_g(t_k)$ (2) can be constructed by calculating the deviation of inter genre once again after calculating the genre representative term.

2.2 Other classification methods

Two algorithms that are applied to evaluate DCM are SVM(Support Vector Machine) and Bayesian model. SVM was proposed by Vapnik and it identifies the vector based on the hyperplane between two groups then it finds the maximized margin in the hyperplane. SVM is used in many applications[8,9] and it is well known for excellent accuracy[10].

The Bayesian model[11] is an algorithm that can be estimated by multiplying the probability of a term appearing in specific class andthe ratio of the number of documents in the specific class to the total number of documents in the training set. Bayesian model is used in many applications[2,12] because it has comparatively simple processing and consistent accuracy among many classification models.

2.3 Testing environment

Three classification models were used to evaluate the classification accuracy. DCM and Bayesian classifiers were implemented with C++ language in windows 7 OS. libSVM[13] by Chang & Lin is applied for SVM.

3. Classification accuracy by the number of training documents

3.1 Construction of the documents set

Brown corpus has been used for some of existing genre classification experiments.

It contains well organized genres but the total

<Table 1> Reconstructed corpus after division

Corpus of Lee & Myaeng			Corpus divided into 7 sets						
Genre	Subject classes within each genre	n	A	B	C	D	E	F	G
Reportage	*robbery,accident,fraud,killing,violence,drug,kidnap,suicide, ..*	930	133	133	133	133	133	133	132
Editorial	*economy,education,international,culture,north Korea,society,*	750	107	108	107	107	107	107	107
Thesis	*engineering,education,basic science,biomedical,agriculture,..*	1,050	150	149	150	150	150	151	150
Review	*education,finance,culture,sports,shopping,cloths,computers,..*	2,234	319	320	319	319	319	319	319
Blogs	*students,teachers,professors,celebrity,employee, ..*	906	129	129	129	129	130	130	130
Q&A	*laws,customers,English,cuisine,medical,computer,childcare, ..*	960	138	137	137	137	137	137	137
Spec	*jewely,sports,video,computer,cosmetics,motors,cell phone, ..*	870	124	124	125	125	124	124	124
Total		7,700	1,100	1,100	1,100	1,100	1,100	1,100	1,100

〈Table 2〉 Organization of the training documents

Testing docs # of training docs		A	B	C	D	E	F	G
level1 ~ level9	21,28,35,50, 100,200,300, 600,900	extracted from B	extracted from A	extracted from A	extracted from A	extracted from A	extracted from A	extracted from A
level10	1,100	B	A	A	A	A	A	A
level11	2,200	B,C	A,C	A,B	A,B	A,B	A,B	A,B
level12	3,300	B,C,D	A,C,D	A,B,D	A,B,C	A,B,C	A,B,C	A,B,C
level13	4,400	B,C,D,E	A,C,D,E	A,B,D,E	A,B,C,E	A,B,C,D	A,B,C,D	A,B,C,D
level14	5,500	B,C,D,E,F	A,C,D,E,F	A,B,D,E,F	A,B,C,E,F	A,B,C,D,F	A,B,C,D,E	A,B,C,D,E
level15	6,600	B,C,D,E,F,G	A,C,D,E,F,G	A,B,D,E,F,G	A,B,C,E,F,G	A,B,C,D,F,G	A,B,C,D,E,G	A,B,C,D,E,F

〈Table 3〉 Testing B by DCM after training it with 21 docs from set A

Genre Testing set B	Reportage	Editorial	Thesis	Review	Blogs	Q&A	Spec	Precision/ Recall
Reportage(108)	78	26	3	1	0	0	0	0.716/0.722
Editorial(133)	18	115	0	0	0	0	0	0.701/0.865
Review(320)	0	5	314	1	0	0	0	0.844/0.981
Thesis(149)	1	2	0	145	0	0	1	0.810/0.973
Blogs(129)	7	10	16	22	65	3	6	0.985/0.504
Q&A(137)	2	5	27	8	1	89	5	0.967/0.650
Spec(124)	3	1	12	2	0	0	106	0.898/0.855
Classification accuracy(micro averaging)								0.829

number of containing documents is around 500 which is too small for this study. Corpus[6,7] of Lee & Myaeng contains about 7,700 documents which are enough to evaluate the classification accuracy with reliable objectivity. Reorganized corpus of Lee & Myaeng is used in this study.

Corpus of Lee & Myaeng consists of 7 genres (Reportage, Editorial, Thesis, Review, Blogs, Q&A and Spec) that are preferred by internet users. Table 1 shows that the corpus is equally divided into 7 sets to optimize the evaluation for objective result. Each set from A to G has 1,100 documents in total that contains 7 subsets of each genre. Set A, for example, has 133 Reportage documents, 107 Editorials, 150 Theses, 319 Reviews, 129 Blogs, 138 Q&As and 124 Specifications.

Table 2 shows that the number of training documents for each setis divided into 15 levels. For example, 21 training documents are extracted from set B for level 1 for testing document set A and 6,600 training documents are extracted from set B to set G for level 15 for testing document set A. This helps to evaluate the performance of each classification model by comparing the accuracy with the changing the number of training documents. The training documents from level 1 to level 9 are extracted from document set A (except testing document set A which uses training documents that are extracted from documents set B). It is possible to diversify the components of training documents, but the number of cases is minimized to achieve more concise results and to obtain greater degrees of reliability and objectivity. Training documents level 1,2 and 3 are all multiples of 7 so that they can include 3,4 or 5 documents from each genre.

3.2 Classification accuracy for the levels of training

Table 3 shows the result from analyzing set B by DCM after training it with 21 documents that are extracted from set A.

Looking at the first row, we can see that among the 108 documents in the ‘reportage’genre, 78 documents are correctly assigned and 26 are incorrectly assigned to the ‘editorial’ genre. At the second row, 18 documents among 133 in the ‘editorial’genre are incorrectly assigned to the ‘reportage’ genre. These errors occur because both categories are derived from ‘newspaper’ genre and the fact that they have many common features leads to a loss of discrimination capability.

Total classification accuracy in each experiment is calculated with micro-averaging to complement the relatively different numbers of testing documents in each genre.

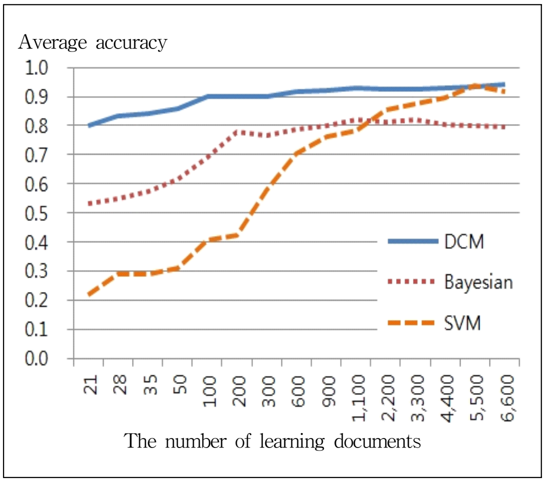
〈Table 4〉 Classification accuracy by DCM with 21 training docs

Training documents	Testing documents	Classification accuracy
21 extracted from B	set A	0.909
21 extracted from A	set B	0.829
21 extracted from A	set C	0.795
21 extracted from A	set D	0.760
21 extracted from A	set E	0.764
21 extracted from A	set F	0.766
21 extracted from A	set G	0.779
Average accuracy		0.800

Table 4 shows classification accuracy that was gained from analyzing 7 documents set A to G with DCM after training them with 21 documents. The shaded cell in table 4 shows the classification accuracy of table 3.

Table 5 shows average classification accuracy from the experiment with 15 levels of training documents by DCM, Bayesian and SVM. The shaded cell shows the average classification accuracy of table 4.

Figure 1 shows the data converted table 5 into a graph chart. (Figure 1 shows the relationship between the average accuracy and the different number of documents in Table 5.) DCM showed rapid increase of accuracy when it had up to 100 training documents. The accuracy increased gradually to 1,100 training documents and it did not show any significant improved accuracy with the increased number of documents.



[Fig. 1] Graph chart of the average accuracy

Bayesian showed a rapid increase of accuracy up to 200 documents. It showed increase and decrease repeatedly between 200 and 3,300 training documents and the accuracy went up to 0.822. Above 3,300 documents, the accuracy begins to decline.

SVM showed a low accuracy of less than 0.6 with up to 300 training documents. With a small number of learning documents SVM showed difficulty in adapting to new documents which are beyond the former margin of the hyperplane.

SVM showed a gradual improved accuracy with an increase of training documents. When it had 5,500 training documents, it showed accuracy of 0.936 which is even higher than the accuracy of 0.933 by DCM, but SVM showed a decreased accuracy with 6,600 training documents. SVM showed a proportional increase in accuracy with a changing number of training documents. After it reaches the limit number of training documents, it doesn’t show further improvement in accuracy.

〈Table 5〉 Classification accuracy with the changing number of training documents

Test sets # of training docs			A	B	C	D	E	F	G	Average accuracy
L1	21	DCM	0.909	0.829	0.795	0.760	0.764	0.766	0.779	0.800
		Bayesian	0.635	0.581	0.540	0.522	0.478	0.496	0.466	0.531
		SVM	0.097	0.284	0.267	0.228	0.230	0.248	0.165	0.217
L2	28	DCM	0.933	0.860	0.820	0.803	0.814	0.788	0.803	0.832
		Bayesian	0.651	0.624	0.544	0.545	0.512	0.505	0.462	0.549
		SVM	0.290	0.292	0.290	0.290	0.291	0.295	0.291	0.291
L3	35	DCM	0.932	0.873	0.826	0.805	0.819	0.809	0.830	0.842
		Bayesian	0.710	0.637	0.561	0.554	0.535	0.525	0.488	0.573
		SVM	0.290	0.291	0.290	0.290	0.290	0.290	0.290	0.290
L4	50	DCM	0.942	0.865	0.849	0.840	0.841	0.826	0.849	0.859
		Bayesian	0.763	0.648	0.632	0.587	0.562	0.564	0.543	0.614
		SVM	0.290	0.292	0.336	0.389	0.291	0.295	0.291	0.312
L5	100	DCM	0.953	0.931	0.901	0.876	0.872	0.886	0.878	0.900
		Bayesian	0.774	0.739	0.697	0.679	0.633	0.634	0.669	0.689
		SVM	0.471	0.424	0.405	0.444	0.212	0.435	0.447	0.405
L6	200	DCM	0.950	0.922	0.895	0.876	0.876	0.880	0.898	0.900
		Bayesian	0.842	0.829	0.773	0.778	0.768	0.706	0.741	0.777
		SVM	0.681	0.460	0.435	0.386	0.329	0.340	0.344	0.425
L7	300	DCM	0.935	0.922	0.903	0.882	0.874	0.877	0.904	0.900
		Bayesian	0.833	0.790	0.777	0.740	0.747	0.713	0.761	0.766
		SVM	0.809	0.664	0.610	0.517	0.482	0.506	0.455	0.578
L8	600	DCM	0.933	0.946	0.924	0.907	0.892	0.906	0.905	0.916
		Bayesian	0.809	0.849	0.778	0.776	0.756	0.768	0.782	0.788
		SVM	0.920	0.784	0.729	0.732	0.578	0.580	0.599	0.703
L9	900	DCM	0.935	0.956	0.935	0.915	0.903	0.908	0.900	0.922
		Bayesian	0.868	0.845	0.815	0.791	0.780	0.734	0.776	0.801
		SVM	0.896	0.847	0.783	0.798	0.645	0.645	0.718	0.762
L10	1,100	DCM	0.940	0.962	0.936	0.915	0.901	0.915	0.930	0.928
		Bayesian	0.891	0.854	0.827	0.830	0.790	0.743	0.817	0.821
		SVM	0.905	0.873	0.809	0.816	0.690	0.657	0.735	0.784
L11	2,200	DCM	0.955	0.969	0.939	0.911	0.895	0.898	0.900	0.924
		Bayesian	0.848	0.895	0.853	0.804	0.794	0.752	0.753	0.814
		SVM	0.953	0.939	0.868	0.864	0.765	0.785	0.807	0.854
L12	3,300	DCM	0.959	0.964	0.940	0.939	0.887	0.887	0.898	0.925
		Bayesian	0.849	0.887	0.857	0.880	0.804	0.755	0.727	0.822
		SVM	0.959	0.950	0.895	0.887	0.799	0.807	0.836	0.876
L13	4,400	DCM	0.955	0.962	0.959	0.939	0.896	0.877	0.905	0.928
		Bayesian	0.821	0.875	0.869	0.827	0.775	0.714	0.759	0.806
		SVM	0.972	0.954	0.920	0.960	0.820	0.784	0.846	0.894
L14	5,500	DCM	0.954	0.969	0.961	0.958	0.907	0.877	0.905	0.933
		Bayesian	0.806	0.876	0.853	0.844	0.767	0.704	0.760	0.801
		SVM	0.970	0.952	0.962	0.945	0.862	0.935	0.925	0.936
L15	6,600	DCM	0.962	0.962	0.959	0.964	0.915	0.910	0.915	0.941
		Bayesian	0.838	0.850	0.839	0.850	0.721	0.742	0.732	0.796
		SVM	0.976	0.970	0.954	0.943	0.882	0.833	0.877	0.919

The following clarifies the performances of the three classification methods at around the accuracy of 0.8, DCM showed 0.800 with 21 training documents of level 1, Bayesian showed 0.801 with 900 training documents of level 9 and SVM showed 0.784 with 1,100 training documents of level 10. This indicates that DCM had a reliable classification accuracy even with small number of training documents when they are organized in subject classes within genres, but Bayesian and SVM shows only if they have enough number of training documents.

4. Conclusion

Classification accuracy is the most critical factor that is used to evaluate classification models. Classification accuracy is affected by the number of learning documents, and it is generally expected that a large number of training documents will bring a high classification accuracy. There are a few studies that show how this actually influences automatic text classification. This study focused on measuring the performance of DCM which is developed for genre-based classification. The goal of the experiments was to measure the classification capability of DCM with a changing number of training documents so as to compare and contrast it with other classification algorithms.

Classification experiments were performed to evaluate the classification accuracy with web documents that are collected and reorganized to 15 levels which contain training documents sample sizes from 21 to 6,600. DCM had excellent feature selection ability even with a small number of documents such as 21 in level 1, only if it was given subject categories within genres. SVM showed low adaptability to documents that are off the boundary of existing hyperplane when it had a small number of documents. The accuracy increased proportionally as the number of learning documents increased. Bayesian also showed a proportional relationship between the accuracy and the number of training documents in a wider perspective, but it is unreliable due to its fluctuating accuracy.

Further research will seek to identify the influences of the number of features, type of features, the number of categories and the quality of training documents on classification accuracy.

REFERENCES

[1] M. Santini, M. Rosso, Testing a Genre-Enabled Application: A Preliminary Assessment, In Proceedings of the 2nd BCS-IRSG Symposium on Future Directions in Information Access, pp. 54-63, 2008.

[2] Z. Zhang, Q. Ye, Z. Zhang, Y. Li, Sentiment Classification of Internet Restaurant Reviews Written in Cantonese, Expert Systems with Applications, Vol.38, No.6, pp. 7674 - 7682, 2011.

[3] A. Fuxman, A. Kannan, A. B. Goldberg, R. Agrawal, P. Tsaparas, J. Shafer, Improving Classification Accuracy Using Automatically Extracted Training Data, In Proceedings of the 15th ACM SIGKDD, pp. 1145-1154, 2009.

[4] Y. Maeda, H. Yoshida, T. Matsushima, Document Classification Method with Small Training Data, In Proceedings of the ICCAS-SICE, pp. 138-141, 2009.

[5] C. Apte, F. Damerau, S. M. Weiss, Automated Learning of Decision Rules for Text Categorization, ACM Transactions on Information Systems, Vol.12, No.3, pp. 233-251, 1994.

[6] Y. B. Lee, S. H. Myaeng, Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization, In Proceedings of the 37th HICSS, 2004.

[7] Y. B. Lee, S. H. Myaeng, Text Genre Classification with Genre-Revealing and Subject-Revealing Features, In Proceeding of the 25th ACM SIGIR, pp. 145-150, 2002.

[8] Y. Li, C. Chen, Research on the Feature Selection Techniques Used in Text Classification, In Proceedings of the 9thFuzzy Systems and Knowledge Discovery(FSKD), pp. 725-729, 2012.

[9] X. Zhang, W. Xiao, Clustering based Two-stage Text Classification Requiring Minimal Training Data, In Proceedings of the International Conference on System and Informatics(ICSAI), pp. 2233-2237, 2012.

[10] J. Novovicova. Text Document Classification,

ERCIM News, No.62, pp. 53-54, 2005.

[11] D. Lewis, R. Schapire, J. Callan, R. Papka, Training Algorithms for Linear Text Classifiers, In Proceedings of the 19th ACM SIGIR, pp. 298-306, 1996.

[12] R. Jayashree, K. Srikantamurthy, S. A. Basavaraj, Suitability of Naïve Bayesian Methods for Paragraph Level Text Classification in the Kannada Language Using Dimensionality Reduction Technique, International Journal of Artificial Intelligence & Applications(IJAIA), Vol.4, No.5, pp. 121-131, 2013.

[13] C. C. Chang, C. J. Lin, LIBSVM Tools, http://csie.ntu.edu.tw/~cjlin/libsvmtools/

이용배(Lee, Yong Bae)

- 1998년 2월 : 충남대학교 컴퓨터과학과(이학석사)
- 2003년 2월 : 충남대학교 컴퓨터과학과(이학박사)
- 2003년 9월 ~ 현재 : 전주교육대학교 컴퓨터교육과 교수
- 관심분야 : 정보검색, 컴퓨터교육
- E-Mail : yblee@jnue.kr