문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구

Improving the Performance of a Fast Text Classifier with Document-side Feature Selection

이 재 윤*
Jae-Yun Lee

차 례

초 록

문서분류에 있어서 분류속도의 향상이 중요한 연구과제가 되고 있다. 최근 개발된 자질값투표 기법은 문서자동분류 문제에 대해서 매우 빠른 속도를 가졌지만, 분류정확도는 만족스럽지 못하다. 이 논문에서는 새로운 자질선정 기법인 문서측 자질선정 기법을 제안하고, 이를 자질값투표 기법에 적용해 보았다. 문서측 자질선정은 일반적인 분류자질선정과 달리 학습집단이 아닌 분류대상 문서의 자질 중 일부만을 선택하여 분류에 이용하는 방식이다. 문서측 자질선정을 적용한 실험에서는, 간단하고 빠른 자질값투표 분류기로 SVM 분류기만큼 좋은 성능을 얻을 수 있었다.

키 워 드

문서자동분류, 자질선정, 자질가중치, 자질값투표

* 경기대학교 문헌정보학과 전임강사
(Full-time lecturer, Library & Information Science Dept., Kyonggi Univ., memexlee@kyonggi.ac.kr)
• 논문접수일자 : 2005년 11월 18일
• 게재확정일자 : 2005년 12월 9일

ABSTRACT

High-speed classification method becomes an important research issue in text categorization systems. A fast text categorization technique, named feature value voting, is introduced recently on the text categorization problems. But the classification accuracy of this technique is not good as its classification speed. We present a novel approach for feature selection, named document-side feature selection, and apply it to feature value voting method. In this approach, there is no feature selection process in learning phase; but real-time feature selection is executed in classification phase. Our results show that feature value voting with document-side feature selection can allow fast and accurate text classification system, which seems to be competitive in classification performance with Support Vector Machines, the state-of-the-art text categorization algorithms

KEYWORDS

Text Categorization, Feature Selection, Feature Weight, Feature Value Voting

1. 서 론

문서자동분류에 대한 연구는 정보검색과 마찬가지로 1960년대에 시작되었으나, 본격화된 것은 다양한 기계학습이론이 도입된 이후부터라고 볼 수 있다. 의사결정 트리(Lewis and Ringuette 1994), 나이브베이즈 분류기(McCallum and Nigam 1998), *k*NN(*k* Nearest Neighbor)(Masand et al. 1992), 신경망(Wiener et al. 1995), 그리고 최근에 각광받고 있는 SVM(Support Vector Machine)(Joachims 1998)에 이르기까지 문서분류에 적용된 기계학습 알고리듬은 모두 나름대로의 성과를 보여왔다.

그러나 문서분류라는 과제는 전통적으로 기계학습이론이 개발·적용되어온 영역과는 구별되는 특징을 가지고 있으므로 실용성 면에서 아직도 해결해야할 여지가 남아있다.

무엇보다도 분류대상 문서를 표현하는 자질(feature)로 쓰이는 색인어가 매우 다양하다는 것이 문서자동분류의 실용화를 방해하는 주요 원인이라고 할 수 있다. 문서의 양이 조금만 늘더라도 기계학습을 위해 표현하는데 사용되는 색인어 수는 지수함수적으로 증가하기 때문이다. 기계학습 분야에서 자주 언급되는 '차원의 저주(curse of dimensionality)' (Bellman 1961)라는 표현이 가장 잘 어울리는 영역이 문서자동분류이다.

'차원의 저주'를 극복하기 위한 접근방법으로는 상대적으로 자질수에 영향을 덜 받는 알

고리듬을 적용하는 것과 자질차원을 축소하는 방법을 들 수 있다. 일반적으로 나이브베이즈와 같이 자질수에 큰 영향을 받지 않는 고속 분류기는, 문서분류 문제에서 SVM이나 kNN분류기보다 성능이 낮은 것으로 알려져 있다.

이 연구에서는 최근에 제안된 새로운 분류 알고리듬인 자질값투표 기법을 보완할 수 있는 새로운 자질선정 방식을 사용함으로써 빠르면서도 정확한 문서분류 시스템을 구현하고자 하였다.

2. 자질값투표 기법

Deng et al.(2002)은 분류자질 선정기준으로 주로 사용되어온 로그 승산비(log odds ratio)를 분류단서로 이용하는 간단한 문서분류기를 제안하면서 승산비 기반 문서분류(odds ratio based text classification)라고 불렀다. 이들이 제안한 방식은 분류대상 문서에 포함된 각 자질과 개별 후보 범주 사이의 학습된 연관성을 적합성점수(relevance score)로 규정하고, 이 적합성점수를 범주별로 합산하여 가장 큰 값을 가지는 범주로 문서를 할당하는 것이다. 적합성점수로는 학습과정에서 미리 산출되는 로그 승산비를 사용하기 때문에 문서를 분류할 때에는 출현한 자질의 범주별 적합성점수를 단순히 합산하면 된다.

Deng et al.(2002)의 실험결과는 SVM 분류기보다 좋은 경우와 나쁜 경우로 엇갈리게 나타났다. 학습집단에서 분류범주의 크기를 동일하게 조작한 실험에서는 승산비 기반 분류가 좋은 성능을 보였고, 분류범주의 크기를 다양하게 구성한 실험에서는 SVM 분류기가 더 좋게 나타났다. 실제 현실에서는 동일한 크기의 분류범주란 존재하기 어려우므로 이들의 실험결과는 별다른 반향을 불러일으키지 못했다(<표 1> 참조).

이 연구에서는 Deng et al.(2002)이 제안한 방식을 일반화하여 자질값투표 분류기(Feature Value Voting Classifier: FV 분류기)라고 부르기로 한다. 여기서 자질값이란 미리 학습된 분류자질과 분류범주간의 연관도를 뜻한다. 자질값으로는 로그 승산비뿐만 아니라 흔히 자질선정을 위한 기준값으로 사용되어온 카이 제곱통계량, GSS 계수, 상호정보량 등이 모두 사용될 수 있다. 연관도를 자질값으로 삼기 위해서는 학습집단에 속한 문서들에 대해서 자질 f_i의 출현 여부와 범주 c_j에 소속 여부를

<표 1> Deng et al.(2002)의 자동분류 실험결과

실험집단	Reuters_100 (범주별 크기 동일)		newsgroup_1 (범주별 크기 다양)		newsgroup_2 (범주별 크기 다양)	
분류기	SVM	FV-LOR	SVM	FV-LOR	SVM	FV-LOR
마이크로 평균 정확률	0.9170	0.8660	0.8490	0.8650	0.8290	0.8360

〈표 2〉 자질과 범주간 2×2 분할표

	범주 c_j 소속	c_j 이외 범주 소속
자질 f_i 출현	a	b
자질 f_i 미출현	c	d

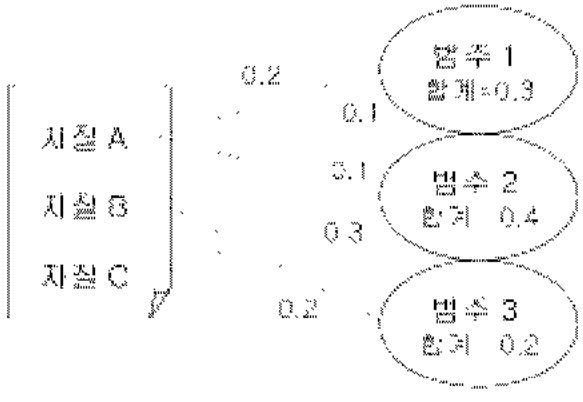
기준으로 2×2 분할표를 〈표 2〉와 같이 구성한 다음 특정 연관성척도 공식을 적용하여 산출한다.

분류대상 문서에 나타난 n개의 단어자질집합과 후보범주 m개의 집합을 각각 $F=\{f_1, f_2, \cdots, f_n\}$와 $C=\{c_1, c_2, \cdots, c_m\}$로 표현하고, 자질 f_i가 범주 c_j에 대해서 가지는 자질값을 $V(f_i, c_j)$라고 하면 자질값투표 분류기는 다음 공식을 만족하는 범주 c_j를 문서에 할당한다.

$$\underset{c_j \in C}{\operatorname{argmax}} \sum_i V(f_i, c_j)$$

즉, 자질값투표 분류기에서는 분류대상 문서에 나타난 각 단어자질과의 연관도 합계가 가장 큰 범주가 문서에 할당된다. 대상문서에 포함된 자질들이 각 범주에 대해서 일종의 가치투표(value voting)를 하여 적합범주를 결정하는 셈이므로 자질값투표(feature value voting)라는 명칭을 사용하였다.

예를 들어, 〈그림 1〉에서와 같이, 분류 대상 문서에 출현한 자질 A, B, C가 학습을 통해 후보범주 1, 2, 3에 대해서 각자 (0.2, 0.1, 0.0), (0.1, 0.0, 0.2), (0,0, 0.3, 0.0)의 연관도를 가지고 있다고 가정해 보자. 이 경우에 각 자질이 가진 연관도에 따라 가치투표하면 범주 1은 0.3(자질A 0.2 + 자질B 0.1), 범주 2는 0.4(자질A 0.1 + 자질C 0.3), 범주 3은 0.2(자질B 0.2)의 값을 가져서 합계가 가장 큰 범주 2가 문서에 할당되게 된다.



〈그림 1〉 자질값투표 분류의 예

이와 같은 자질값투표 분류기는 시간복잡도가 O(mn)으로 나이브베이즈 분류기만큼 단순하다. 더군다나 확률을 곱하는 나이브베이즈 분류기와 달리 자질값을 더하는 방식이기 때문에 미출현 범주에서의 자질값을 0이 아닌 값으로 평활화할 필요가 없으므로 더 빠른 분류가 가능하다.

결국 자질값투표 분류기의 과제는 빠른 분류속도에 비해서 얼마나 높은 분류성능을 얻을 수 있는가라고 할 수 있다. 자질값투표 분류기에 적합한 자질선정 방식과 자질값 산출기준을 찾아내서 자질값투표 분류기의 분류성능을 높이는 것이 이 연구의 목표이다.

3. 자질값투표 분류기에 적합한 자질 선정 방식

3.1 분류자질 선정기준과 자질값 산출기준의 일치

승산비 기반 분류를 제안한 Deng et al.(2002)은 정보획득량(Information Gain)을 자질선정 기준으로 이용하여 SVM 분류기와 성능을 비교해 보았다. 이들의 실험결과는 자질선정이 승산비 기반 분류의 성능을 거의 향상시키지 못하는 것으로 나타났다. 세 실험집단 중 두 실험집단에서는 모든 자질을 사용하는 경우가, 나머지 한 집단에서도 80% 자질을 사용하는 경우가 가장 분류성능이 좋았다. 반면에 SVM 분류기는 자질집단을 10% 내지 20%로 대폭 축소한 경우가 가장 좋아서 상반된 결과가 나타났다.

이들의 실험에서 문제점으로 지적할 수 있는 부분은 분류자질 선정기준과 자질값 산출기준의 불일치이다. 정보획득량과 로그 승산비는 매우 이질적인 척도로서 정보획득량은 고빈도 자질을, 로그 승산비는 저빈도 자질을 우선적으로 채택한다고 알려져 있다(이재윤 2005). Deng et al.(2002)의 실험에서는 자질의 값으로 저빈도 자질을 높게 쳐주는 로그 승산비를 사용하면서도 정작 중요한 자질을 선정할 때에는 저빈도 자질을 배제하는 성질을 가진 정보획득량을 기준으로 사용하였다.

자질값으로서 정보획득량과 로그 승산비가 실제로 어느 정도로 다른 값인가를 알아보기 위해서, 이 연구에서 사용할 실험집단 중 하나인 TREND 2287 실험집단에 포함된 7,544개의 단어 자질에 대해서 문헌빈도, 출현한 범주의 수, 정보획득량, 로그 승산비를 각각 산출한 다음 이들 간의 스피어맨 상관계수를 측정하였다.

<표 3>에 나타난 값을 보면 정보획득량은 문헌빈도와 0.803, 출현한 범주수와 0.413으로 양의 상관관계를 가진 것으로 나타난다. 반면에 로그 승산비는 문헌빈도와 0.371, 출현한 범주수와 0.701로 음의 상관관계를 가진 것으로 나타난다. 이는 정보획득량은 고빈도 자질일수록 높은 반면에, 로그 승산비는 고빈도 자질일수록 낮은 값을 가진다는 것을 뜻한다.

이를 감안하면, 자질의 중요성을 로그 승산

〈표 3〉 자질값으로서 정보획득량과 로그 승산비의 상관관계 (Spearman's rho)

		문헌빈도	출현범주수	정보획득량	로그 승산비
문헌빈도	Correlation Coefficient	1.000	0.667	0.803	-0.371
	Sig. (2-tailed)	.	0.000	0.000	0.000
	N	7544	7544	7544	7544
출현범주수	Correlation Coefficient	0.667	1.000	0.413	-0.701
	Sig. (2-tailed)	0.000	.	0.000	0.000
	N	7544	7544	7544	7544
정보획득량	Correlation Coefficient	0.803	0.413	1.000	0.078
	Sig. (2-tailed)	0.000	0.000	.	0.000
	N	7544	7544	7544	7544
로그 승산비	Correlation Coefficient	-0.371	-0.701	0.078	1.000
	Sig. (2-tailed)	0.000	0.000	0.000	.
	N	7544	7544	7544	7544

비로 산출하여 자질값으로 사용할 경우에는 일관되게 자질선정 기준으로도 로그 승산비를 사용하는 것이 합리적이라고 추측된다. 또한 로그 승산비가 아닌 다른 연관성척도를 이용한 자질값투표 분류기를 구현하여 성능을 비교할 경우에도, 자질값 산출기준과 자질선정 기준을 동일하게 적용해야 특정 척도가 자질의 중요성을 판단하는 능력에 대한 공정한 비교가 될 것이다.

따라서 이 연구에서 검증하고자 하는 〈가설 1〉은 다음과 같다.

〈가설 1〉 : 자질값투표 분류기에서 자질값 산출 기준인 로그 승산비를 자질선정 기준으로도 사용하면 다른 기준을 자질선정 기준으로 사용하는 것보다 좋은 성능을 얻을 수 있을 것이다.

〈가설 1〉의 검증을 위해서 분류자질 선정기준과 자질값 산출기준을 일치시킨 경우와 그렇지 않은 경우의 성능을 비교하는 〈실험 1〉을 5.1절에서 수행하였다.

3.2 문서측 자질선정

일반적인 분류자질선정은 사전 학습단계에서 학습문서집단을 대상으로 분류자질을 선정한 다음 분류정보를 학습한다. 이 연구에서 제안하는 문서측 자질선정(document side feature selection)은 이와 달리 학습단계에서 자질을 선정하지 않고 분류정보를 학습한 다음, 분류실행단계에서 분류대상 문서의 자질 중 일정한 수 또는 일정한 비율만큼을 선정하여 분류에 이용하는 방식이다. 따라서 동일한 자질이 문서에 따라서 선정될 수도 있고, 제외될 수도 있다. 기존의 분류자질 선정방식을 (학습)집단측(collection side) 자질선정, 오프라

인 자질선정, 학습 시 자질선정이라고 부른다면, 제안하는 방식은 (대상)문서측(document side) 자질선정, 온라인 자질선정, 분류 시 자질선정이라고 부를 수 있다.

문서측 자질선정이 필요한 이유는, 학습집단측 자질선정을 적용할 경우에는 개별 문서를 고려하지 않고 전체 학습집단 차원에서 분류자질을 선정하므로 실제 분류대상 문서를 표현할 때에 사용되는 자질의 수가 문서마다 매우 다르기 때문이다. 사전 실험을 통해 확인해본 결과 학습집단측 자질선정을 수행하면 분류대상 문서를 표현할 분류자질의 수가 어떤 문서는 2개밖에 안되는가 하면 어떤 문서는 70 내지 80개가 되는 것으로 나타났다. 너무 적은 경우는 분류에 실패할 가능성이 높고, 너무 많은 경우는 불필요한 처리시간이 추가로 들게 된다. 문서측 자질선정을 통해서 일정 수의 자질만 남도록 보장하면 이런 문제를 해결할 수 있을 것이다. 이를 반영한 실험 〈가설 2〉는 다음과 같다.

〈가설 2〉 : 자질값투표 분류기에서 문서측 자질선정을 할 경우에, 너무 적은 수의 자질만 사용하면 낮은 성능을 얻을 것이며, 또한 일정한 수 이상의 자질사용은 성능향상에 도움이 되지 않을 것이다.

적절한 수의 문서측 자질선정이 분류성능향상에 도움이 될 것이라는 〈가설 2〉를 검증하기 위해서 5.2절의 〈실험 2〉를 수행하였다. 〈실험 2〉에서는 문서측 자질선정 실험을 수행하면서 문서당 자질의 수를 1개에서부터 90개 이상까지 늘려가면서 분류성능의 변화를 관찰하여 적정성능을 보이는 문서당 자질수를 파악하였다.

4. 실험 설계

4.1 실험문서집단과 실험용 시스템

이 연구에서는 〈표 4〉와 같이 두 가지 분류실험용 문서집단을 이용하였다.

〈표 4〉 실험에 사용된 문서집단

실험문서집단	KFCM-896	TREND-2287
내용	신문기사	해외과학기술문헌속보
문서의 수 [전체 / 학습 / 검증]	[896 / 718 / 178]	[2,287 / 1,178 / 1,109]
범주의 수	17	8
범주별 학습문서 수 [평균 / 최대 / 최소]	[42.2 / 81 / 17]	[147.3 / 424 / 27]
(저빈도어 제거 후) 학습집단의 색인어 종수	7,261	7,544
(저빈도어 제거 후) 학습문서의 평균 색인어 종수	88.3	97.9
(저빈도어 제거 후) 학습문서의 평균 색인어 수	151.4	161.8

KFCM 896 분류실험집단은 1992년 신문기사로 구성된 KFCM CL1020(정영미, 이재윤 2001) 실험집단에서 주요 대분류 항목인 정치, 경제, 산업, 국제 분야에 속한 기사 896건만 추출한 것이다.

TREND 2287 분류실험집단은 정보검색용 실험집단인 HANTEC v.2.0(김지영 외 2000)에서 분류정보가 포함된 해외과학기술문헌속보 문서 2,287건을 추출한 것이다. 2,287건 중에서 1997년 4/4분기 3개월간 등록된 문서 1,178건을 학습문서집단으로, 1998년 1/4분기 3개월간 등록된 문서 1,109건을 검증문서집단으로 이용하였다. 문서의 분류는 현재 해외과학기술동향 홈페이지(http://techtrend.kisti.re.kr/)에서 구분한 9개 대분류 중 '과학기술 일반'을 제외한 8개 대분류를 적용하였다.

각 문서는 제목과 본문을 대상으로 자동색인하였고 규모가 작은 KFCM 896은 추출된 색인어 중에서 장서빈도(CF)가 2 이하인 경우를, 이보다 규모가 큰 TREND 2287은 문헌빈도(DF)가 2 이하인 경우를 전처리 단계에서 제거하였다.

자질값투표 분류기는 Visual FoxPro로 구현하였고 성능비교를 위한 SVM 분류기는 WEKA version 3.4(Witten and Frank 2005)를 사용하였다.

4.2 세부실험 단계

〈실험 1〉에서는 〈가설 1〉의 검증을 위해서 자질값투표 분류기와 SVM 분류기의 성능을 학습집단측 자질선정과 함께 비교해본다. 구체적인 내용은 다음과 같다.

① 로그 승산비(LOR)를 자질값으로 사용한 자질값투표 분류기(이하 FV LOR 분류기로 표기함)의 성능을 SVM 분류기와 비교한다.

② Deng et al.(2002)과 마찬가지로 정보획득량을 기준으로 학습집단측 자질선정을 수행하였을 때의 FV LOR 분류기와 SVM 분류기의 성능을 비교한다.

③ FV LOR 분류기에 대해서 학습집단측 자질선정을 하되 자질선정 기준을 자질값 산출기준과 같은 로그 승산비로 일치시켰을 때의 성능을 알아본다.

이를 통해서 자질값투표 분류기의 성능을 재확인하고, 자질값투표 분류기에 적합한 학습집단측 자질선정 방식을 확인하고자 한다.

〈실험 2〉에서는 〈가설 2〉의 검증을 위해서 로그 승산비를 이용한 자질값투표 분류기 FV LOR에 대해서 문서측 자질선정의 효과를 검증한다. 문서측 자질선정에서는 분류대상 개별문서에 포함된 자질 중에서 범주별 최대 자질값이 높은 f개의 자질만을 선정하여 문서표현에 사용한다. 범주별 최대 자질값이란 한 자질이 각 범주에 대해서 가지는 자질값 중에서 가장 큰 것을 말한다. 이때 선정된 자질의 자질값을 그냥 합산하는 기본 방식과 자질값에 문서내 가중치를 곱해서 합산하는 변형방식을 함께 비교해본다.

〈실험 3〉에서는 로그 승산비 이외의 다른 연

〈표 5〉 자질값 산출기준으로 검토한 연관성척도

연관성척도	약호	빈도수준 선호경향	공식 (〈표 2〉의 2×2 분할표에 적용)
GSS	GSS	고빈도 선호	$\frac{ad-bc}{N^2}$
코사인 (Ochiai라고도 함)	COS	고빈도 선호	$\frac{a}{\sqrt{(a+b)(a+c)}}$
피어슨 상관계수 (Pearson's PHI라고도 함)	PCC	중간빈도 선호	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
상대적 상호정보량 J	RMIJ	중저빈도 선호	$\frac{\log_2 N+\log_2 a-\log_2(a+b)(a+c)}{\log_2 N-\log_2 a}$
로그 승산비	LOR	저빈도 선호	$\log\frac{ad}{bc}$
율의 Y	YULE	저빈도 선호	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
상호정보량	MI	저빈도 선호	$\log_2\frac{Na}{(a+b)(a+c)}$

관성척도를 자질값투표 분류기에서 이용해본다. 추가로 검토한 척도는 연관성척도의 빈도수준 선호경향(이재윤 2004)에 따라서 〈표 5〉와 같이 성향이 다양한 척도를 포함시켰다. 이 실험을 통해서 자질값투표 분류기에 어울리는 연관성척도의 유형을 알아볼 수 있다.

각 실험의 성능평가는 분류시도 건수 대비 분류성공 건수로 산출하는 마이크로 평균 정확률 척도를 기준으로 하였다.

5. 실험결과 분석

5.1 〈실험 1〉 – 학습집단측 자질선정 적용

두 실험집단 KFCM 896과 TREND 2287에 대해서 SVM 분류기와 FV LOR 분류기를 사용하여 학습집단측 자질선정을 적용하면서 분류한 결과를 실험집단별로 〈표 6〉과 〈그림 2〉, 그리고 〈표 7〉과 〈그림 3〉에 제시하였다.

〈표 6〉과 〈그림 2〉는 KFCM 896 실험집단을 대상으로 수행한 실험결과이다. 여기서는 정보획득량을 기준으로 자질선정하였을 때 SVM 분류기의 최고성능은 40% 자질집합을 사용한 0.5955였고, FV LOR 분류기의 최고성능은 이보다 낮은 0.5674로 나타났다. 반면에 로그 승산비를 기준으로 자질선정한 FV LOR 분류기의 최고성능은 0.6348로 월등히 높게 나타났다.

〈표 7〉과 〈그림 3〉은 TREND 2287 실험집단을 대상으로 수행한 실험결과이다. 정보획득

〈표 6〉 SVM과 FV-LOR 분류기 성능비교 (KFCM-896)

자질집합 크기	SVM (IG 기준 자질선정)	FV-LOR (IG 기준 자질선정)	FV-LOR (LOR 기준 자질선정)
10%	0.5730	0.5225	0.4270
20%	0.5787	0.5337	0.4719
30%	0.5843	0.5506	0.5674
40%	0.5955	0.5674	0.5955
50%	0.5899	0.5618	0.6011
60%	0.5730	0.5618	0.5955
70%	0.5618	0.5393	0.6348
80%	0.5730	0.5674	0.6236
90%	0.5618	0.5674	0.6180
100%	0.5618	0.5674	0.5674

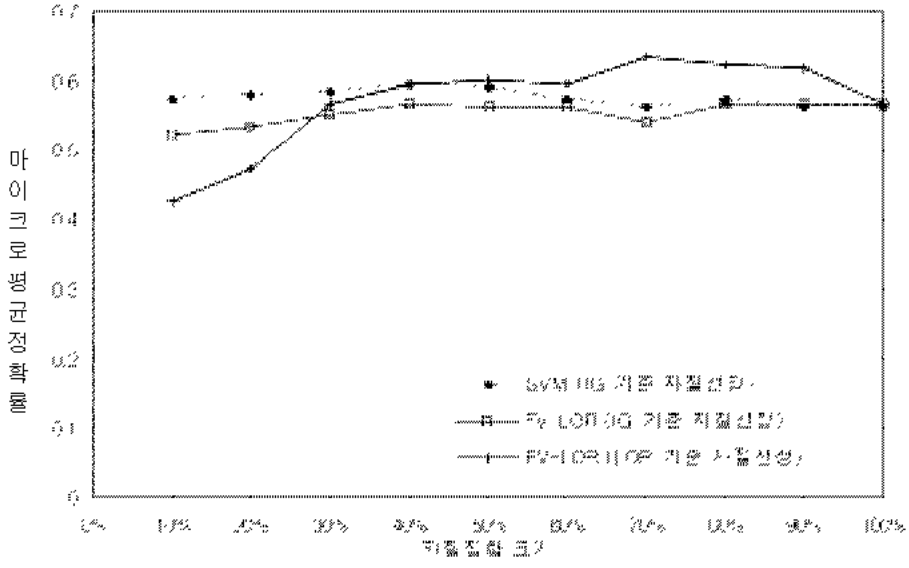
*음영 부분은 해당 분류기에서 성능이 가장 좋은 경우

〈표 7〉 SVM과 FV-LOR 분류기 성능비교 (TREND-2287)

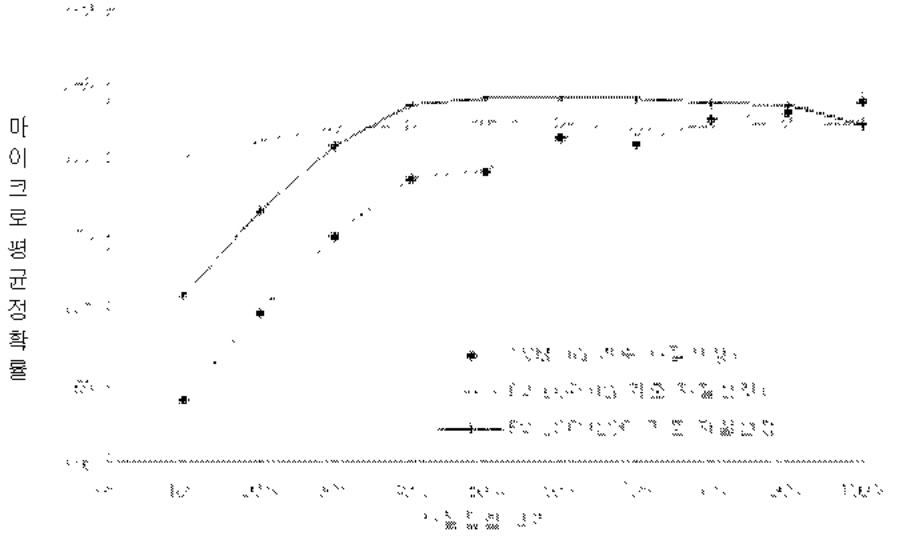
자질집합 크기	SVM (IG 기준 자질선정)	FV-LOR (IG 기준 자질선정)	FV-LOR (LOR 기준 자질선정)
10%	0.6411	0.8025	0.7106
20%	0.6979	0.8124	0.7665
30%	0.7493	0.8215	0.8097
40%	0.7872	0.8224	0.8368
50%	0.7926	0.8260	0.8404
60%	0.8151	0.8251	0.8413
70%	0.8106	0.8197	0.8404
80%	0.8269	0.8224	0.8377
90%	0.8314	0.8251	0.8368
100%	0.8386	0.8242	0.8242

* 음영 부분은 해당 분류기에서 성능이 가장 좋은 경우

량을 기준으로 자질선정하였을 때 SVM 분류기의 최고성능은 자질선정을 적용하지 않은 100% 자질집합을 사용한 0.8386이었고, FV LOR 분류기의 최고성능은 50% 자질집합을 사



〈그림 2〉 SVM과 FV-LOR 분류기 성능비교 (KFCM-896)



〈그림 3〉 SVM과 FV-LOR 분류기 성능비교 (TREND-2287)

용한 0.8260으로 SVM의 성능보다 더 낮게 나타났다. 반면에 로그 승산비를 기준으로 자질선정한 FV LOR 분류기의 최고성능은 60% 자질집합을 사용하였을 때의 0.8413으로 나타나서, 정보획득량으로 자질선정한 경우의 FV LOR 분류기는 물론 SVM 분류기보다도 좋은

성능을 보였다.

〈실험 1〉을 통해 얻어진 결과는 다음과 같다.

첫째, 자질선정을 하지 않은 경우에 KFCM 896 집단에서는 FV LOR 분류기(0.5674)가 SVM 분류기(0.5618)보다 약간 성능이 좋았으며, TREND 2287 집단에서는 SVM 분류기가 더 좋았다. 학습집단에서 분류범주의 크기 편차가 더 심한 TREND 2287 집단에서 FV LOR 분류기가 나쁜 성능을 보인 것은 Deng et al.(2002)의 실험결과와 일맥상통하는 부분이다.

둘째, 정보획득량(IG)을 기준으로 자질선정을 한 경우에 일부 실험집단에서는 SVM 분류기가 자질선정을 통해 성능을 향상시킬 수 있었던 반면에 FV LOR 분류기는 성능향상 효과를 거의 얻지 못하였다. KFCM 896 집단에서는 SVM 분류기가 3.37% 포인트의 뚜렷한 성능향상을 보인 반면에 FV LOR 분류기는 성능이 향상되지 않았다. TREND 2287 집단에서는 SVM 분류기도 자질선정으로 성능이 향상되지 않았으며 FV LOR 분류기는 약간 향상되었으나 차이는 0.18% 포인트로 미미하였다. 이것 역시 Deng et al.(2002)의 실험과 유사한 결과라고 할 수 있다.

셋째, FV LOR 분류기에 대해서 로그 승산비를 기준으로 자질선정한 경우에는 두 실험집단 모두에서 뚜렷하게 성능이 향상되어 SVM 분류기보다 더 좋은 성능을 보였다. 이는 FV 분류기에서 자질선정 기준과 자질값 산출기준이 상반되지 않도록 해야 함을 뜻한다.

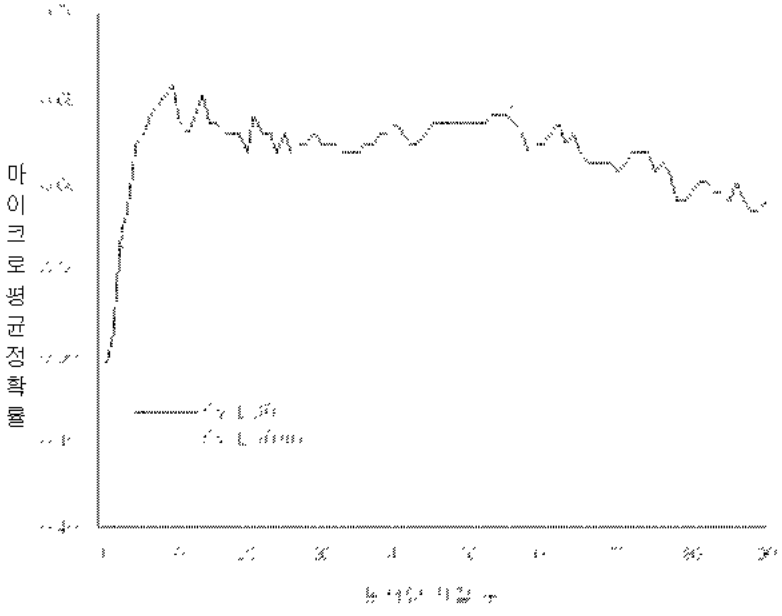
이로써 자질값투표 분류기에서 자질값 산출기준인 로그 승산비를 자질선정 기준으로도 사용하면 다른 기준을 자질선정 기준으로 사용하는 것보다 좋은 성능을 얻을 수 있을 것이라는 〈가설 1〉은 검증되었다.

5.2 〈실험 2〉 – 문서측 자질선정 적용

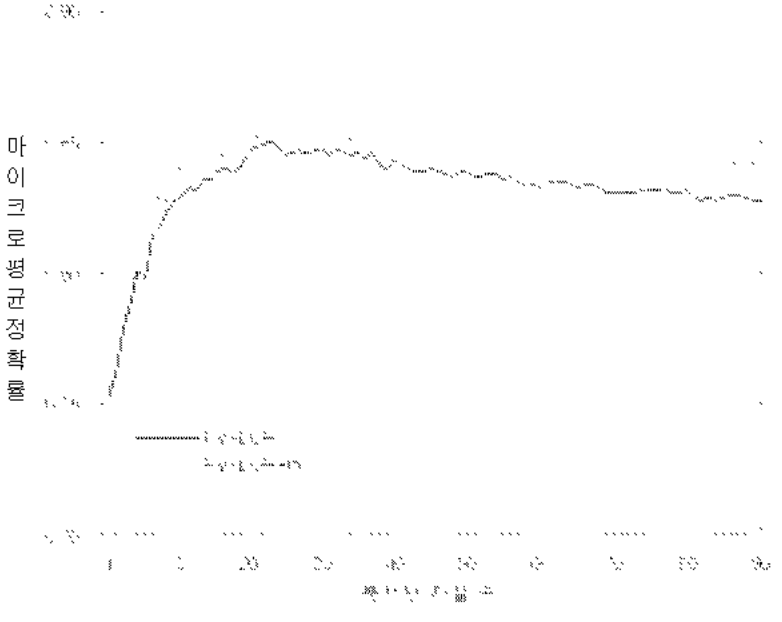
〈실험 2〉에서는 FV LOR 분류기에서 각 분류대상 문서마다 대표자질값이 높은 f개의 자질만 선정해서 분류에 적용해보았다. 선정된 자질의 자질값을 그냥 합산하는 기본 방식과 자질값에 문서 내 가중치(로그 TF 공식으로 산출한 ln 가중치)를 곱해서 합산하는 변형방식(FV LOR * ln으로 표기)을 함께 비교해본다. 두 실험집단에서 분류대상 문서당 자질수 f를 1개에서 90개까지 늘려가며 실험한 결과는 각각 〈그림 4〉, 〈그림 5〉과 같다.

〈그림 4〉와 〈그림 5〉에서는 문서당 자질의 수를 1개에서부터 10여개까지 추가할 때에는 성능이 급격하게 향상된다. 이는 문서당 자질수가 너무 적을 경우에는 분류성능이 낮게 나타날 것이라는 〈가설 2〉를 확인해주는 결과이다.

또한, 문서측 자질선정 결과 문서당 자질의 수를 10개에서 15개 정도 이상 사용하면 안정적인 성능에 달하는 것으로 보인다. 문서길이가 짧은 KFCM 896 집단에서는 10개, 문서길이가 이보다 긴 TREND 2287 집단에서는 15개 정도 이상이면 된다. 그 이상으로 자질수를 늘리게 되면 경우에 따라서는 오히려 성능이



〈그림 4〉 문서측 자질선정 결과 (KFCM-896)



〈그림 5〉 문서측 자질선정 결과 (TREND-2287)

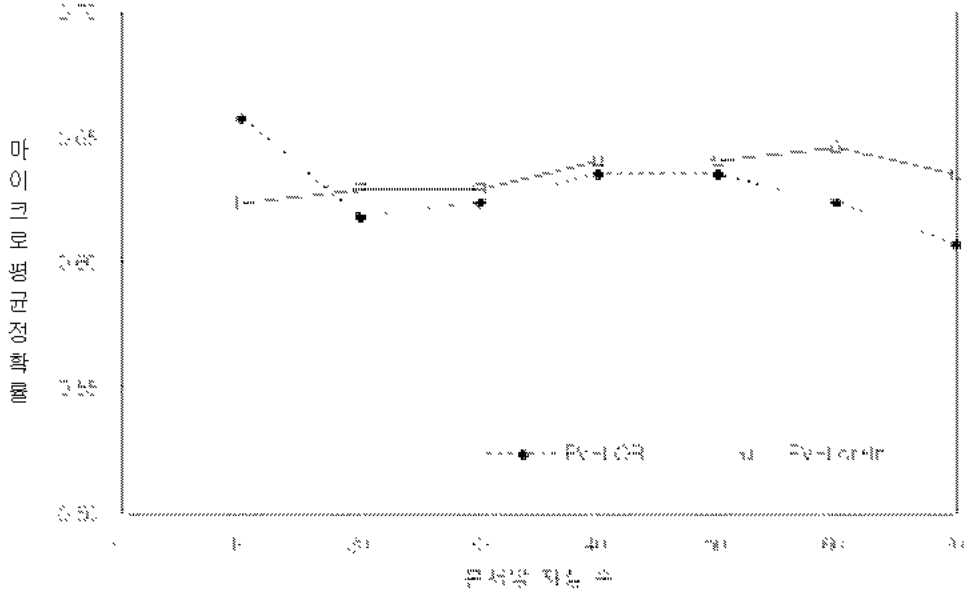
약간 저하되는 현상도 나타난다. 이것 역시 문서당 자질수를 일정한 수 이상으로 사용하는 것은 분류성능을 향상시키지 못할 것이라는 〈가설 2〉의 주장을 뒷받침하는 결과이다.

문서당 자질수 f를 10개에서 70개까지 10개씩 늘렸을 경우의 성능을 수치로 살펴보면 〈표 8〉, 〈그림 6〉, 〈그림 7〉과 같다.

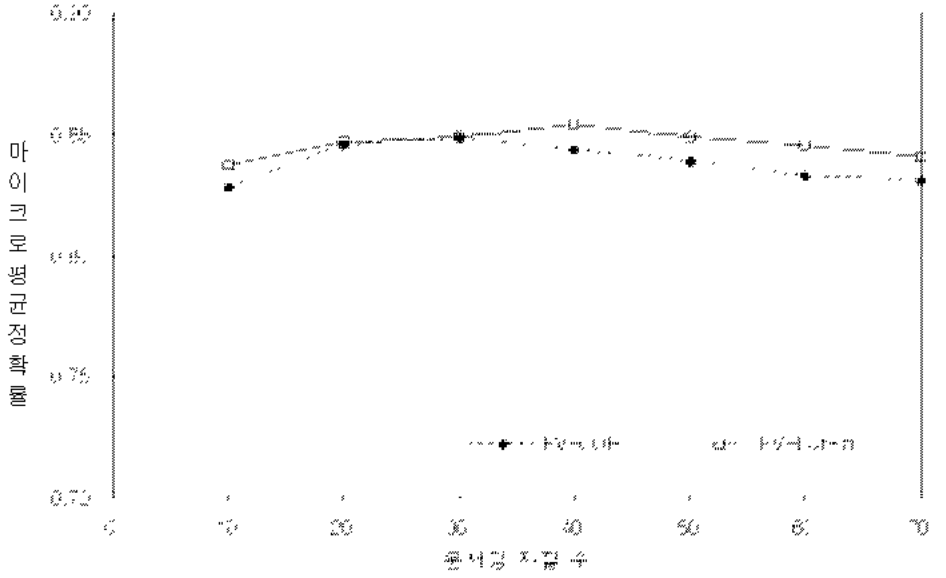
〈표 8〉에 나타난 문서측 자질선정을 적용한

〈표 8〉 문서측 자질선정 결과

문서당 자질수	KFCM-896		TREND-2287	
	FV-LOR	FV-LOR*In	FV-LOR	FV-LOR*In
10	0.6573	0.6236	0.8278	0.8368
20	0.6180	0.6292	0.8449	0.8467
30	0.6236	0.6292	0.8476	0.8485
40	0.6348	0.6404	0.8431	0.8539
50	0.6348	0.6404	0.8386	0.8485
60	0.6236	0.6461	0.8323	0.8449
70	0.6067	0.6348	0.8305	0.8404



〈그림 6〉 문서측 자질선정 결과 (KFCM-896 실험집단)



〈그림 7〉 문서측 자질선정 결과 (TREND-2287 실험집단)

성능을 〈표 6〉, 〈표 7〉의 학습집단측 자질선정을 적용한 FV LOR 분류기의 성능과 비교해 보면 최고성능의 경우에, KFCM 896 실험집단에서는 0.6348에서 0.6573으로 3.54% 높게 나타났다. 또한 TREND 2287 실험집단에서는 0.8413에서 0.8539로 1.50% 향상된 성능을 얻었다. 특히 분류대상 문서 내 가중치를 이용한 FV LOR * ln 방식은 〈그림 6〉, 〈그림 7〉에서 볼 수 있듯이 최고성능이 높을 뿐만 아니라 문서당 자질수를 변화시키더라도 상당히 안정적인 성능을 얻을 수 있는 것으로 나타났다.

5.3 〈실험 3〉 - 로그 승산비 이외의 연관성척도 적용

〈실험 3〉에서는 로그 승산비 이외에 고빈도, 중간빈도, 저빈도 수준을 선호하는 연관성척도를 각각 두 가지씩 채택하여 FV 분류기에 적용해보았다. 자질선정을 하지 않은 경우(FV), 학습집단측 자질선정을 적용한 경우(FV+CSFS), 문서측 자질선정을 적용한 경우(FV+DSFS), 문서측 자질선정을 하면서 자질값에 문서 내 가중치를 곱해서 합산한 경우(FV+DSFS(ln))의 네 가지로 나누어 각각 최고성능을 구한 결과를 〈표 9〉과 〈표 10〉, 〈그림 8〉과 〈그림 9〉에 제시하였다. 여기서 자질선정 기준으로는 자질값 산출기준과 동일한 척도를 사용하였다.

학습집단측 자질선정에서는 10% 포인트씩 자질집합의 크기를 줄여나간 결과 중 가장 좋은 성능을, 문서측 자질선정에서는 f 값을 10에서 70까지 10개씩 늘이면서 실험한 결과 중 가장 좋은 성능을 실험결과로 제시하였다.

〈표 9〉 연관성척도에 따른 자질값투표 분류기의 분류성능 비교 (KFCM-896)

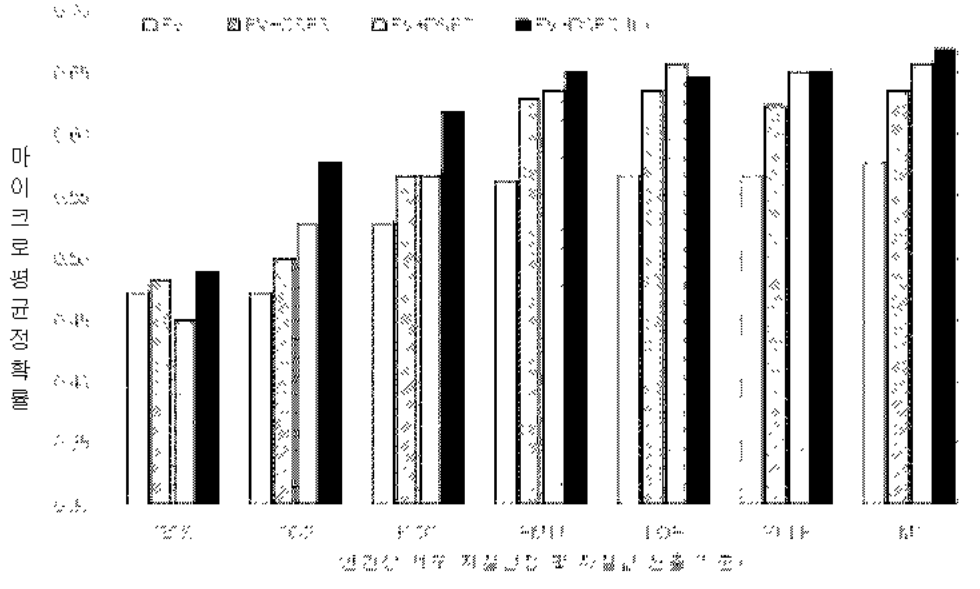
	GSS	COS	PCC	RMIJ	LOR	YULE	MI
FV	0.4719(-20.8%)	0.4719(-20.8%)	0.5281(-11.3%)	0.5618(-5.7%)	0.5674(-4.7%)	0.5674(-4.7%)	0.5787(-2.8%)
FV+CSFS	0.4832(-18.9%)	0.5000(-16.0%)	0.5674(-4.7%)	0.6292(+5.7%)	0.6348(+6.6%)	0.6236(+4.7%)	0.6348(+6.6%)
FV+DSFS	0.4494(-24.5%)	0.5281(-11.3%)	0.5674(-4.7%)	0.6348(+6.6%)	0.6573(+10.4%)	0.6517(+9.4%)	0.6573(+10.4%)
FV+DSFS(ln)	0.4888(-17.9%)	0.5787(-2.8%)	0.6180(+3.8%)	0.6517(+9.4%)	0.6461(+8.5%)	0.6517(+9.4%)	0.6685(+12.3%)

* 괄호 안은 SVM 분류기의 최고성능(0.5955)과 비교한 향상률

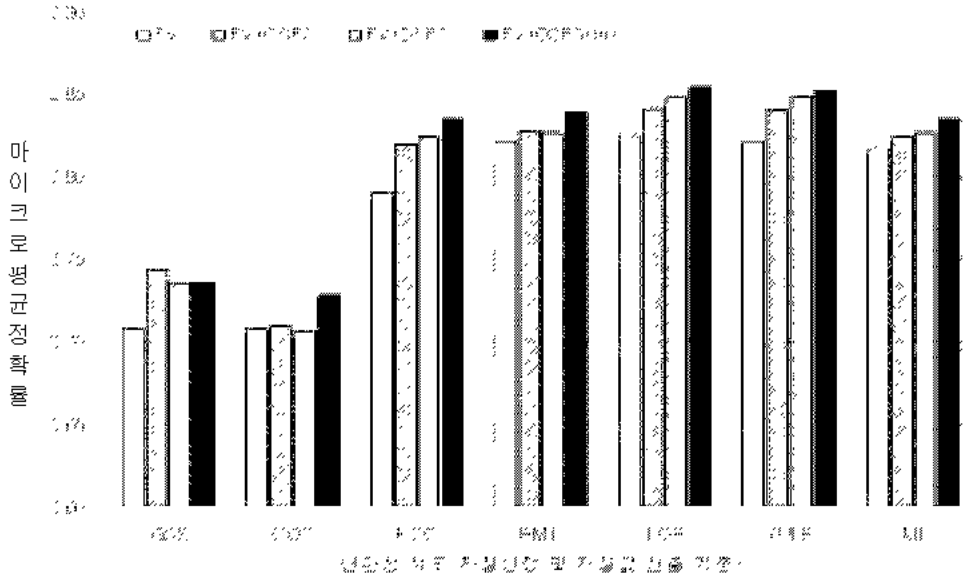
〈표 10〉 연관성척도에 따른 자질값투표 분류기의 분류성능 비교 (TREND-2287)

	GSS	COS	PCC	RMIJ	LOR	YULE	MI
FV	0.7069(-15.7%)	0.7069(-15.7%)	0.7899(-5.8%)	0.8206(-2.2%)	0.8242(-1.7%)	0.8206(-2.2%)	0.8161(-2.7%)
FV+CSFS	0.7439(-11.3%)	0.7088(-15.5%)	0.8188(-2.4%)	0.8269(-1.4%)	0.8413(+0.3%)	0.8404(+0.2%)	0.8233(-1.8%)
FV+DSFS	0.7350(-12.3%)	0.7060(-15.8%)	0.8233(-1.8%)	0.8260(-1.5%)	0.8476(+1.1%)	0.8485(+1.2%)	0.8260(-1.5%)
FV+DSFS(ln)	0.7358(-12.3%)	0.7286(-13.1%)	0.8350(-0.4%)	0.8386(+0.0%)	0.8539(+1.8%)	0.8521(+1.6%)	0.8350(-0.4%)

* 괄호 안은 SVM 분류기의 최고성능(0.8386)과 비교한 향상률



〈그림 8〉 연관성척도별 분류성능비교 (KFCM-896)



〈그림 9〉 연관성척도별 분류성능비교 (TREND-2287)

〈실험 3〉의 결과는 다음과 같다.

첫째, 자질선정을 하지 않은 성능은 자질값 산출기준으로 사용한 연관성척도가 저빈도선호 경향을 가질수록 좋게 나타났다. 고빈도 자질을 선호하는 GSS 계수와 코사인 계수를 사용했을 때에는 분류기와 자질선정 방식에 상관없이 항상 SVM 분류기보다 낮은 성능을 보였다.

둘째, 자질선정을 한 경우의 성능은 학습집단에 대해서 한 경우(FV+CSFS)보다 문서측에 대해서 한 경우(FV+DSFS)가 좋았고, 자질값을 합산할 때 문서 내 가중치를 곱한 경우(FV+DSFS(ln))가 이보다도 더 좋았다.

셋째, 연관성척도별로는 로그 승산비(LOR)와 율의 Y(YULE)가 가장 좋았고 상호정보량과 상대적 상호정보량 J를 이용한 경우가 그 다음으로 좋은 성능을 보였다.

넷째, SVM 분류기의 성능과 비교하였을 때 두 실험집단에서 모두 더 좋은 성능을 보인 경우는, 로그 승산비와 율의 Y를 자질값 산출기준으로 사용하고 문서측 자질선정을 하되 자질값에 문서 내 가중치를 곱해서 합산한 경우(FV+DSFS(ln))였다.

다섯째, 상대적으로 학습집단의 범주별 크기가 고른 편인 KFCM 896 집단에서는, 자질값투표 분류기와 문서측 자질선정을 적용하였을 때의 성능이 SVM 분류기보다도 좋게 나타난 경우가 많았다.

6. 결 론

자질값투표 분류기는 간단하고 빠르지만 전통적인 고빈도어 위주의 자질선정을 통해서는

좋은 성능을 얻을 수 없었다. 가장 큰 이유는 자질값 산출기준과 자질선정 기준 간의 불일치 때문이었으며, 또한 학습집단측 자질선정에 따라서 분류대상 문서별로 사용가능한 자질의 수가 심하게 차이난 것도 원인이었다.

이 연구에서 제시한 문서측 자질선정 방식은 일반적인 자질선정 방식처럼 학습집단을 대상으로 학습단계에서 자질선정을 수행하지 않고, 분류대상 문서를 처리할 때 자질선정을 동시에 진행하는 실시간 자질선정 방식이다. 문서측 자질선정 방식은 자질값 산출기준과 자질선정 기준 간의 불일치 문제를 해결함과 동시에, 분류대상 문서별 자질수가 심하게 차이나는 문제를 해결할 수가 있었다. 두 실험집단을 대상으로 이루어진 세 차례에 걸친 실험결과를 요약하면 다음과 같다.

첫째, 전통적인 학습집단측 자질선정에 있어서 자질선정 기준과 자질값 산출기준을 로그 승산비로 일치시킨 결과, 자질값투표 분류기의 성능이 SVM 분류기의 성능과 비슷하거나 더 좋은 것으로 나타났다.

둘째, 문서측 자질선정을 적용한 실험에서는 학습집단측 자질선정의 경우보다 1% 포인트 이상 좋은 성능을 얻었다. 특히 자질값투표를 실행할 때, 미리 학습한 자질값과 그 자질의 분류대상 문서 내 가중치를 곱해서 사용할 경우에는, 분류성능이 더 향상되었을 뿐만 아니라 문서당 자질수를 변화시키더라도 매우 안정적인 성능을 얻을 수 있었다.

셋째, 로그 승산비 이외의 다른 연관성척도를 자질값 산출기준과 자질선정 기준으로 사용해본 결과, 율의 Y를 제외하고는 모두 로그 승산비를 사용한 경우보다 항상 낮은 성능을 보였다. 율의 Y도 공식에 승산비가 포함된다는 점을 감안하면, 자질값투표 분류기에는 승산비를 사용하는 공식이 적합함을 알 수 있다.

문서의 빠른 분류가 가능한 자질값투표 분류기에 적합한 자질선정 방식을 통해서 SVM 분류기 이상의 성능을 얻을 수 있었다. 향후 연구에서는 대량의 실험집단에 대해서 제안된 방식을 검증하고자 한다. 또한, 문서측 자질선정 방식을 자질값투표 분류기 이외의 다른 분류 알고리듬에 적용해보는 연구도 필요하다.

참고문헌

김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. 한국어 테스트 컬렉션 HANTEC의 확장 및 보완. 『제 12회 한글 및 한국어 정보처리 학술대회 논문집』, 210 215.

이재윤. 2004. 연관성척도의 빈도수준 선호경향에 관한 연구. 『정보관리학회지』, 17(4): 281 294.

이재윤. 2005. 자질선정 기준과 가중치 할당 방식간의 관계를 고려한 문서자동분류의 개선에 대한 연구. 『한국문헌정보학회지』, 39(2): 123 146.

정영미, 이재윤. 2001. 지식 분류의 자동화를 위한 클러스터링 모형 연구. 『정보관리

학회지』, 18(2): 203 230.

Bellman, R. E. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Deng, Zhi Hong, Shi Wei Tang, Dong Qing Yang, Ming Zhang, Xiao Bin Wu, and Meng Yang. 2002. "Two odds radio based text classification algorithms." *Proceedings of the Third International Conference on Web Information Systems Engineering* (Workshops), 223 231.

Joachims, T. 1998. "Text categorization with support vector machines: learning with many relevant features." *Proceedings of 10th European Conference on Machine Learning* (ECML 98), 137 142.

Lewis, D. D., and M. Ringuette. 1994. "A comparison of two learning algorithms for text categorization." *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, 81 93.

Masand, G., G. Linoff, and D. Waltz. 1992. "Classifying news stories using memory based reasoning." *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 59 65.

McCallum, A., and K. Nigam. 1998. "A comparison of event models for Naive Bayes text classification." *Proceedings of AAAI 98 Workshop on Learning for Text Categorization*, 41 48.

Wiener, E., J. O. Pedersen, and A. S. Weigend. 1995. "A neural network approach to topic spotting." *Proceedings of Fourth Annual Symposium on Document Analysis and Information Retrieval* (SDAIR' 95), 317 332.

Witten, Ian H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. (2nd ed.). San Francisco: Morgan Kaufmann.