문서의 불균등 분포를 고려한 단어 불순도 기반 특징 선택 방법

(An Enhanced Feature Selection Method Based on the Impurity of Words Considering Unbalanced Distribution of Documents)

강 진 범 † 양 재 영 †† 최 중 민 †††
(Jinbeom Kang) (Jaeyoung Yang) (Joongmin Choi)

요 약 기계 학습 과정에서 수집된 많은 정보들 중에는 학습하고자 하는 개념과 관련이 없거나 중복된 정보를 가진 경우가 많다. 또한 자료 자체에 오류가 있기도 하다. 이와 같이 학습 모델 생성을 위해 수집된 정보를 신뢰할 수 없다면, 학습 과정에서도 정확한 지식 습득이 어렵다. 그래서 기계 학습은 학습 과정에서 정확한 지식 습득을 위해 특징 선택 방법을 사용한다. 특징 선택은 학습할 클래스와 관련이 없거나 중복된 정보를 학습 모델 생성 이전에 제거함으로써 학습 알고리즘의 성능을 향상시킨다. 기존의 특징 선택 방법들은 적절한 특징을 선택하기 위하여 문서가 균등하게 분포되어 있다고 가정한다. 하지만, 실제로는 그렇지 않으며, 문서의 수 또는 문서의 길이가 모두 동일한 학습 예제를 준비하는 것도 매우 어렵다.

본 논문에서는 보다 효율적으로 특징을 선택하기 위해 클래스 별 단어의 불순도와 문서의 불균등 분포를 고려한 특징 선택 방법을 제안한다. 클래스를 대표할 수 있는 특징 후보들을 단어의 불순도 측정을 통해 얻고, 문서의 불균등 분포를 고려하여 특징을 선택한다. 실험을 통해 보다 좋은 성능을 보임을 입증한다.

키워드 : 특징 선택, 기계학습, 문서분류, 단어 불순도, 문서 불균등 분포

Abstract Sample training data for machine learning often contain irrelevant information or redundant concept. It is also the case that the original data may include noise. If the information collected for constructing learning model is not reliable, it is difficult to obtain accurate information. So the system attempts to find relations or regulations between features and categories in the learning phase. The feature selection is to remove irrelevant or redundant information before constructing learning model. for improving its performance. Existing feature selection methods assume that the distribution of documents is balanced in terms of the number of documents for each class and the length of each document. In practice, however, it is difficult not only to prepare a set of documents with almost equal length, but also to define a number of classes with fixed number of document elements.

In this paper, we propose a new feature selection method that considers the impurities among the words and unbalanced distribution of documents in categories. We could obtain feature candidates using the word impurity and eventually select the features through unbalanced distribution of documents. We demonstrate that our method performs better than other existing methods via some experiments.

Key words : feature selection, machine learning, classification, word impurity, unbalanced distribution of documents

· 이 논문은 "국가 IT 온톨로지 인프라 기술개발" 정보통신부 선도과제 성과의 일부입니다.

† 학생회원 : 한양대학교 컴퓨터공학과
jbkang@cse.hanyang.ac.kr
†† 정 회 원 : 동부정보기술 RFID/USN Part Manager
isconan@dongbu.com
††† 종신회원 : 한양대학교 컴퓨터공학과 교수
jmchoi@cse.hanyang.ac.kr

논문접수 : 2006년 8월 21일
심사완료 : 2007년 6월 11일

1. 서 론

기계 학습(machine learning) 분야에서 특정 문서를 대표할 수 있는 단어 및 패턴의 집합을 문서를 나타낼 수 있는 특징(feature)이라 한다. 기계 학습은 두 단계로 이루어지는데, 학습(learning) 단계에서는 특징들과 클래스간의 관계나 규칙성을 찾기 위한 시도를 하고,

분류(classification) 단계에서는 학습 단계에서 습득한 학습 모델을 이용하여 이전에 알지 못했던 새로운 예제(unseen examples)에 대한 클래스를 결정한다.

효과적인 문서 분류를 위해서는 학습하고자 하는 개념과 관련된 많은 특징들이 필요하다. 하지만 수집된 많은 정보들 중에는 학습하고자 하는 개념(concept)과 관련이 없거나 중복된 정보를 가진 경우도 있다. 또한 자료 자체에 오류가 포함되어 있기도 하다. 이와 같은 학습 모델 생성을 위해 수집된 데이타를 신뢰할 수 없다면, 학습 과정에서도 정확한 지식의 습득이 어렵다[1].

특징 선택(feature selection)의 과정은 학습할 개념과 관련이 없거나 중복된 정보를 학습 모델 생성 이전에 제거함으로써 학습 알고리즘의 성능을 향상시키기 위해 학습 알고리즘이 수행되기 전의 전처리 과정으로 사용된다. 이러한 과정을 통해서 많은 자료들 중 실제 분류 성능에 영향을 줄 수 있는 특징을 선별해 낼 수 있다. 또한 학습 모델 생성에 사용될 자료의 수를 줄임으로써 학습 알고리즘이 조금 더 빠르고 효과적으로 동작할 수 있으며 생성된 학습 모델의 크기도 줄일 수 있다.

본 논문에서는 개념을 잘 표현할 수 있는 특징에 대해 탐구하고 향상된 특징 선택 방법을 제안한다. 효율적 특징 선택을 위해 두 가지의 측면을 고려하였다. 첫 번째는 특징이 하나의 개념에만 나타나면 그 특징은 다른 개념의 간섭을 받지 않고 특정 개념을 잘 나타낼 수 있는 좋은 특징이 될 수 있다는 것을 증명하고, 이를 특징 선택에 반영할 수 있는 기법을 탐구하였다. 두 번째는 학습을 하기 위해 수집하는 예제 집합들이 현실적으로 균등하게 분포되지 않는다는 문제점을 파악하고, 이를 보완하기 위한 방법을 제시하였다.

2. 관련 연구

특징 선택은 기계학습, 통계 데이타 마이닝, 패턴 인식 분야에서 활발히 연구되고 있다. 초기에는 특징의 중복 및 관련성이 적은 특징들을 제거하는 것이 목표였다. 이와 같이 예제 집합으로부터 불필요한 특징들을 제거함으로써 학습 모델 생성 시 발생하는 계산 시간이나 많은 자료의 수집 및 관리에 드는 비용을 줄일 수 있었다. 또한 만들어진 학습 모델로부터 생성되는 규칙들을 보다 쉽게 이해할 수 있다.

특징 선택 알고리즘은 크게 wrapper 접근법과 filter 접근법의 두 가지 범주로 분류한다[2,3]. wrapper 접근법은 명확한 특징 평가하기 위해 반복적으로 수행하게 된다. 그래서 속도는 느리지만 성능 면에서 그 유용성이 입증되었다. 하지만 반복적은 수행으로 속도가 느리기 때문에 많은 대량의 자료 집합에서는 잘 사용되지 않는다. 반면 filter 접근법은 학습 알고리즘과는 독립적으로 동작하며, 자료들의 일반적인 성질을 기반으로 선택된 특징 집합을 평가한다. 이 방법은 wrapper 접근법보다 빠르기 때문에 예제 자료 집합이 많이 사용되는 문서에서는 더 효율적이라 볼 수 있다.

분류 알고리즘을 실사회에 존재하는 거대 문서 자료 집합에 적용하기 위해서는 반드시 문서를 표현하는데 이용되는 특징의 수를 축소시켜야 한다[4]. 이를 위해 [5]에서 문서 집합과 특징 집합간의 연관을 평가하여 특징을 추출한 후 ID3, C4.5를 이용해 분류 실험에서 특징들의 차원 축소(dimensionality reduction)가 성능향상에 미치는 효과를 입증하였다. 또한 [4]에서 통계적인 기반의 문서 분류 학습 기법을 이용하여 로이터(Reuters) 데이타에 대해 여러 가지의 특징선택 방법과 k-NN과 LLSF(Linear Least Squares Fit mapping) 알고리즘을 적용한 분류실험에서 원래 문서의 98%를 제거한 특징만을 가지고도 더욱 정확한 분류를 해낼 수 있음을 입증하였다. [6]에서는 Yahoo사이트에 존재하는 계층적인 웹 문서들에 대해 여러 특징 선택 방법들과 나이브 베이지안(Naive Bayesian) 분류 알고리즘을 이용하여 어떤 특징 선택 방법이 좀 더 나은 성능을 내는지를 실험하였다.

기존의 특징 선택 방법에서 사용되는 기법은 크게 문서 빈도수, 상호 정보, 정보 획득, χ^2(카이제곱) 통계량 등으로 구분할 수 있다.

2.1 문서 빈도수(DF: Document Frequency)

문서 빈도수란 어떤 단어가 나타난 문서의 빈도수를 말하는 것으로 학습 문서에서 그 단어가 나타나는 문서의 빈도수를 계산한 후에 일정 빈도수 이상의 단어만을 특징으로 선택하는 기법이다. 이 기법은 출현 문서 빈도수가 적은 단어는 분류기의 성능에 기여하지 못한다는 것을 가정한다. 가장 간단하고 계산량이 적다는 장점이 있으나 "적은 문서 빈도수를 갖는 단어가 정보량이 많다"라는 정보 검색에서 널리 받아들여지는 기본 가정과 대치되어 일반적으로 잘 사용되지는 않는다[7].

2.2 상호 정보(Mutual Information)

상호 정보는 통계적 언어 모델(statistical language model) 문서관리를 위한 분류기를 구현하기 위해 일반적으로 사용되는 이론 및 기법이다[7]. 상호 정보 기법에서는 클래스 c에서의 단어 t의 정보량을 식 (1)과 같이 나타낸다. 이 식이 의미하는 것은 클래스 c에서 단어 t가 많이 출현할수록 t의 정보량은 크다는 것이다.

$$I(t,c)=\log\frac{\Pr(t\wedge c)}{\Pr(t)\times\Pr(c)} \quad (1)$$

$$I(t,c)\approx\log\frac{A\times N}{(A+C)\times(A+B)} \quad (2)$$

식 (2)는 식 (1)의 근사 값을 계산하기 위해 사용한 식으로, 여기서 A는 클래스 c에 속한 문서 중 단어 t를 포함하는 문서의 수이고, B는 클래스 c 이외의 클래스에 속해 있는 문서에서 단어 t가 출현할 빈도이며, C는 클래스 c에서 단어 t를 포함하지 않는 문서의 수이다. 그리고 N은 전체 학습 문서의 수이다. $I(t,c)$는 단어 t와 클래스 c가 서로 독립적이면 0의 값을 가진다. 전체 학습 문서에서의 단어 정보량은 각 클래스에 대한 단어의 정보량을 계산한 후 평균 정보량 혹은 최대 정보량을 구한다.

$$I_{avg}(t)=\sum_{i=1}^{m}\Pr(c_i)I(t,c_i) \quad (3)$$

$$I_{\max}(t)=\max_{i=1}^{m}I(t,c_i)$$

상호 정보의 단점은 식 (4)에서 알 수 있듯이 같은 조건부 확률값($\Pr(t|c)$)을 갖는 단어라도 전체 출현 빈도($\Pr(t)$)가 적은 단어의 상호 정보량이 상대적으로 더 높게 나온다는 것이다.

$$I(t,c)=\log\Pr(t|c)-\log\Pr(t) \quad (4)$$

2.3 정보 획득(Information Gain)

정보 획득은 기계 학습 분야에서 일반적으로 사용되는 기법이다[8]. 이 기법의 특징은 문서에서의 출현 빈도뿐만 아니라 출현하지 않은 빈도까지 고려해서 각 클래스에서의 단어 정보량을 계산한다는 것이다. $\{c_1, c_2, ..., c_m\}$를 클래스 집합이라 할 때 단어 t의 정보 획득량은 식 (5)를 이용해 구할 수 있다.

$$\begin{aligned}G(t)=&-\sum_{i=1}^{m}\Pr(c_i)\log\Pr(c_i)\\&+\Pr(t)\sum_{i=1}^{m}\Pr(c_i|t)\log\Pr(c_i|t)\\&+\Pr(\bar{t})\sum_{i=1}^{m}\Pr(c_i|\bar{t})\log\Pr(c_i|\bar{t})\end{aligned} \quad (5)$$

식 (5)에서 정보 획득은 모든 클래스의 평균값으로 계산된다. 학습 문서가 주어지면 학습 문서에서 나타나는 모든 단어들의 정보 획득량을 계산하여 일정 임계값 이상의 값을 갖는 단어들만이 특징으로 선택된다.

정보 획득 방법은 상호 정보 척도와는 달리, 단어의 출현 빈도를 고려한 상호 간의 정보 평균값과 단어가 출현하지 않은 빈도의 상호 정보 척도의 평균값의 합으로 계산된다. 이러한 특징으로 인해 정보 획득은 상호 정보 척도보다 문서 분류에서 대략적으로 더 좋은 성능을 보인다.

2.4 χ^2 통계량(χ^2 statistics)

χ^2 통계량은 단어 t와 클래스 c와의 의존성(dependency)을 측정하는 것으로서 자유도(degree of freedom) 1인 χ^2 분포와 비교될 수 있다. χ^2 통계량을 계산하는 식은 식 (6)과 같다.

$$\chi^2(t,c)=\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (6)$$

여기서 A는 클래스 c에 속해 있는 문서 중에서 단어 t를 포함하고 있는 문서의 수이고, B는 클래스 c 이외의 클래스에 속해 있는 문서 중에서 단어 t를 포함하고 있는 문서의 수이다. 또한, C는 클래스 c에 속해 있는 문서 중에서 단어 t를 포함하지 않는 문서의 수이고, D는 클래스 c 이외의 클래스에 속해 있는 문서 중에서 단어 t를 가지고 있지 않은 문서의 수이다. 그리고 N은 전체 학습 문서의 수이다.

χ^2 통계량은 단어 t와 클래스 c가 완전히 독립적이면 0의 값을 가진다. 각 클래스에 대한 단어의 정보량을 계산한 후 전체 학습 문서에서의 단어 정보량을 계산하기 위해서 평균 정보량 혹은 최대 정보량을 구한다.

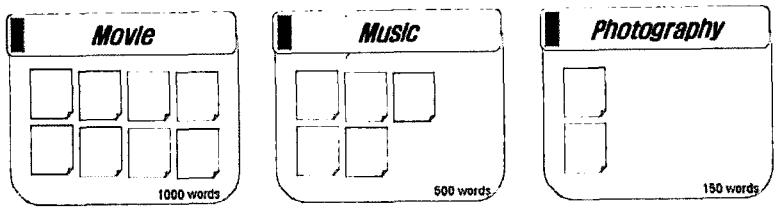
$$\chi^2_{avg}(t)=\sum_{i=1}^{m}\Pr(c_i)\chi^2(t,c_i) \quad (7)$$

$$\chi^2_{\max}(t)=\max_{i=1}^{m}\chi^2(t,c_i)$$

3. 기존 특징 선택 방법의 문제점

기존의 특징 선택 방법은 문서가 균등하게 분포되어 있다는 가정 하에 클래스를 대표할 수 있는 특징 선택에 중점을 두고 있다. 그래서 실제 테스트 작업 시 많은 수의 문서 또는 긴 문서에서 학습된 클래스에 있는 단어가 높은 점수를 얻어 특징 선택의 후보가 될 가능성이 높다. 반대로 적은 수의 문서 또는 짧은 문서에서 학습된 클래스에 있는 단어는 낮은 점수를 얻어 특징으로 선택될 가능성이 낮다.

기존 방법의 이런 단점을 예를 들어 제시하고자 한다. 그림 1과 같이 클래스 3개가 있다고 가정하자. 영화(Movie) 클래스에 속하는 문서는 8개가 존재하고, 음악(Music) 클래스는 5개의 문서, 사진(Photography) 클래스는 2개의 문서를 포함한다. 이 경우 대부분의 기존 방법에서는 클래스 별로 문서에 나타나는 단어들을 정리하여 문서 빈도수(DF) 값에 대한 테이블을 만들게 되고, 그 결과로 그림 1의 아래에 표시된 3개의 테이블과 유사한 구조가 얻어진다. 이것을 모두 모아 단어의 문서 빈도수의 값에 따라 정렬하면 그림 2의 왼쪽 테이블과 같이 되는데, 대부분 그림 2의 오른



Term	DF
Actor	8
Director	7
Hollywood	6
Genre	6
Camera	5

Term	DF
Singer	5
Director	3
Genre	2
Composer	2
Writer	2

Term	DF
Camera	2
Lens	2
Light	2
Film	1
Exposure	1

그림 1 예제 : 클래스에 속한 문서에서의 특징 후보 단어 추출

Term	DF
Director	10
Actor	8
Genre	8
Camera	7
Hollywood	6
Singer	5
Composer	2
Writer	2
Lens	2
Light	2
Film	1
Exposure	1

Term	DF
Actor	8
Hollywood	6
Singer	5
Composer	2
Writer	2
Lens	2

6개의 특징이 선택 되었지만,
특정 클래스에 치우친 형태

그림 2 예제 : 임계값에 의해 선택된 특징들

쪽 그림에서와 같이 정렬된 단어들 중 임계값(threshold)을 정하여 그 값보다 큰 상위 몇 개를 선택해서 특징으로 사용한다. 여기서는 임계값을 2라고 정하였다. (참고로, 단어가 속한 클래스를 그 클래스의 색으로 나타내었다. 즉, 영화 클래스는 갈색, 음악 클래스는 녹색, 사진 클래스는 분홍색으로 구분하였다.)

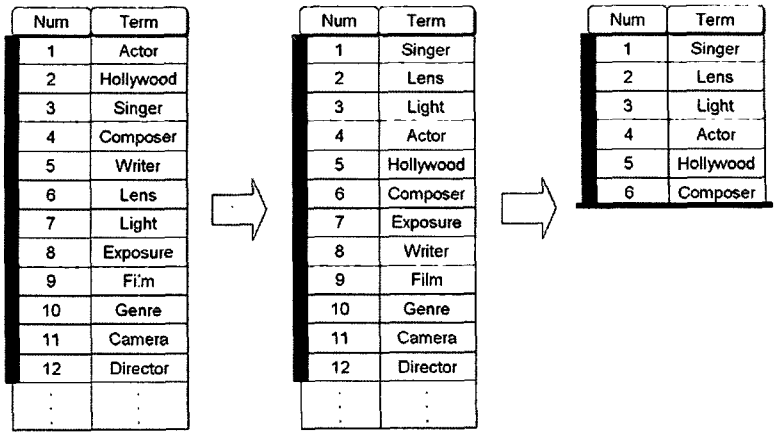
그림 2에서 선택된 특징들 중 상위에 속한 "Director"와 "Genre", "Camera"와 같은 단어는 그림 1에서 보면 하나가 아닌 여러 클래스에서 나타났음을 알 수 있다. (두 가지 이상의 클래스 색깔이 할당되어 있다.) 이 단어들은 특정 클래스 하나만을 가리킬 수 있는 특징들이 되지 못하기 때문에 이런 단어들이 특징으로 선택될 경우 이 단어가 나타난 다른 클래스를 가리킬 수 있다. 이것이 기존 방법의 첫 번째 문제점이 된다. 이에 대한 해결책으로는 다수의 클래스에 나타나는 단어들을 특징 선택 시 배제하는 것이다. 하지만 다수의 클래스에 나타났다고 해서 단순히 배제시키는 것은 비효율적일 수 있으며 이에 대한 체계적 해결책이 요구된다.

위의 예제에서 다수의 클래스에 나타나는 3 단어를 특징 선택에서 제외한다면 그림 2의 오른쪽과 같이 상위 6개의 단어가 특징으로 선택된다. 하지만, 여전히 문서 수와 문서 길이의 값이 상대적으로 큰 영화 클래스에서 나타난 단어들이 높은 점수를 얻고 다음으로 음악 클래스에 속한 단어들, 사진 클래스에 속한 단어들 순으로 정렬되었음을 알 수 있다. 영화 클래스에서 2개, 음악 클래스에서 3개, 사진 클래스에서 1개의 특징이 선택되었다. 이와 같이 클래스 별 문서의 길이 또는 문서 수의 불균등으로 인해 특징이 특정 클래스에 치우친 형태로 선택되었음을 알 수 있다.

하지만, 현실적으로 문서의 수 또는 문서의 길이를 인위적으로 균등하게 하여 학습예제 문서를 준비한다는 것은 어려운 일이다. 만약 균등한 분포 형태로 만들기 위해 억지로 문서 수를 맞추게 되면 학습 자료가 잘 수집된 클래스라 할지라도 학습 문서의 수가 줄게 된다. 줄어든 학습 문서는 학습 양 부족으로 올바르게 동작하지 않을 수 있다. 이러한 문제점은 클래스별 문서의 보유 비율이 다른 점을 가중치로 부여하지 않음으로 인해 발생한다. 따라서 본 논문에서는 이에 대한 가중치 부여 기법을 통해 이를 해결하고자 한다.

4. 단어의 불순도 및 불균등 분포를 고려한 특징 선택 방법

어떤 특징이 제 역할을 하기 위해서는 그 특징이 특정 클래스를 대표할 수 있어야 한다. 이런 특징의 예로는 여러 클래스에 널리 분포되어 있지 않고 오직



(a) 순수도 측정 (b) 불균등 가중치 (c) 특징선택

그림 3 순수도를 고려한 특징 선택

하나의 클래스에서만 나타난 단어들이다. 만일 여러 다른 클래스에서 빈번히 나타난 단어가 특징으로 선택되었다고 가정하자. 임의의 문서를 분류하고자할 때 여러 클래스에 대한 이 문서의 분류 점수가 같아져 잘못된 클래스로 분류될 가능성이 높다. 하나의 클래스에만 나타난 단어인지, 아니면 다수의 클래스에 나타난 단어인지를 구별하기 위해 본 논문에서는 식 (8)과 같은 단어의 불순도(impurity) 측정식을 제시한다.

$$Impurity(c,w) = \Pr(c|w)(\Pr(\bar{c}|w)) \quad (8)$$

이 식은 문서의 특정 단어 w가 클래스 c에 나타날 확률과 나타나지 않을 확률을 곱함으로써 그 단어의 클래스에 대한 불순도를 측정할 수 있다는 것을 의미한다. 상반되는 두 확률 값은 역 관계로, 한 쪽이 커지면, 다른 한 쪽 값은 반대로 작아지는 특성을 가지고 있다. 그래서 두 확률 값은 모두 0이 되거나 1이 될 수 없으며 두 수의 곱은 0과 0.25사이의 값으로 불순도를 측정 할 수 있다. 두 확률 값은 상반되는 역관계이기 때문에 나타날 확률이 1이라면 나타나지 않을 확률 값은 0이 되고 곱은 0이 된다. 이 식은 특정 단어가 다른 클래스에도 널리 나타난 단어인지 판단할 수 있다. 예를 들어 단어가 특정 클래스에만 나타났다고 한다면, 특정 클래스에서 나타날 확률은 1이 되고 다른 클래스에서 나타날 확률은 0이 된다. 따라서 두 수의 곱은 0이 되고 오직 하나의 클래스에 나타났다는 것을 의미한다. 불순도는 얼마나 다른 클래스에도 널리 나타난 단어인지 측정하는 수식이다. 그래서 불순도 값이 커진다면 다른 클래스에도 널리 나타난 단어를 뜻하며, 불순도가 0이면 오직 하나의 클래스에서만 나타난 단어임을 의미한다. 다시 말해 0에 가까운 값을 가지는 단어 w가 클래스 c에서 중요한 역할을 하는 특징임을 의미한다. 역으로 순수도(purity)를 측정하기 위해서는 식 9에서와 같이 1의 값에서 식 (8)의 불순도 값을 뺀다. 단어 w가 클래스 c가 아닌 다른 곳에서 나타날 확률에서 클래스 c에서 나타날 확률의 제곱을 더함으로써 순수도를 얻을 수 있다.

$$\begin{aligned} Purity(c,w) &= 1-\Pr(c|w)\Pr(\bar{c}|w) \\ &= 1-\Pr(c|w)(1-\Pr(c|w)) \\ &= 1-\Pr(c|w)+\Pr(c|w)^2 \\ &= \Pr(\bar{c}|w)+\Pr(c|w)^2 \end{aligned} \quad (9)$$

그림 3의 "(a) 순수도 측정"에서는 여러 클래스에 나타난 정보는 낮은 점수를 가지고 오직 하나의 클래스에만 나타난 단어들이 높은 점수를 얻었다. 하지만, 문서가 많았던 영화 클래스(갈색)에 대한 단어들이 상위권에 있고, 학습 문서가 적었던 사진 클래스(분홍색)에 대한 단어들은 낮은 점수를 얻어 아래에 위치함을 알 수 있다. 이것은 문서 길이와 문서 수의 불균등 분포에 따른 결과이다.

이 문제점을 해결하기 위해 순수도 값에 가중치를 부여하였다. 구체적으로는 식 (10)과 같이 특정 클래스 c에서 단어 w가 나타날 확률 값을 가중치로 부여하였다. $\Pr(w|c)$는 클래스 c에서 단어 w가 어느 정도의 비중을 가지고 있는지를 의미한다.

$$Purity_{weight}(c,w) = Purity(c,w)\Pr(w|c) \quad (10)$$

가중치를 고려한 식을 적용한 결과 그림 3의 "(b) 불균등 가중치"에서와 같이 적은 문서를 학습한 사진 클래스와 음악 클래스의 단어들이 높은 점수를 얻어 "(c) 특징 선택"에서와 같이 클래스 별로 균등하게 특징을 선택할 수 있게 되었다.

클래스 별 단어의 순수도를 측정하여 단어 w가 어느 클래스에서 중요한 역할을 하는지를 알 수 있다.

특징 선택은 두 가지 방법으로 구분된다. 첫 번째는 모든 클래스에 전체적으로 널리 잘 사용될 수 있는 단어를 선택하는 방법으로서 식 (11)의 $PW_{avg}(t)$와 같이 평균값을 구하는 것이고, 두 번째는 특정 클래스에서 좋은 역할을 할 수 있는 단어를 선택하는 방법으로서 식 (11)의 $PW_{max}(t)$와 같이 최대값을 구하는 것이다.

$$PW_{avg}(t) = \sum_{i=1}^{m} \Pr(c_i) Purity_{weight}(c,w) \qquad (11)$$

$$PW_{\max}(t) = \max_{i=1}^{m} Purity_{weight}(c,w)$$

기존의 실험을 통해 최대값을 취하는 방법이 좋다는 것을 알 수 있다[7]. 이것은 하나의 클래스라도 대표할 수 있는 단어가 특징으로 선택되어야 함을 의미한다. 따라서 식 (12)와 같이 순수도, 가중치, 최대값을 고려하는 식이 최종적인 특징선택을 위한 식이 된다.

$$PWM(t) = PW_{\max}(t) \qquad (12)$$

PWM에 의해서 균등하게 단어들이 선택되는 것을 그림 3의 "(c) 특징 선택"을 통해 알 수 있다. 이 예제에서는 각 클래스마다 2개씩의 특징으로 선택되었으며, 이를 통해 문서를 올바른 클래스로 분류할 수 있었다.

5. 실험 및 성능 평가

본 논문에서 제안한 특징 선택 방법의 성능 평가를 위해서 Reuters-21578 자료 집합[9]과 NIPS Conference Papers Vols0-12 raw data[10] 집합에 대해 10-폴드 교차 타당성(10-fold cross validation)[8]으로 평가를 하였다. Reuters-21578 자료 집합은 119개의 클래스와 21578개의 문서로 구성되어 있으며 문서가 매우 불균등하게 분포되어 있다. 예를 들어, 'can' 클래스는 문서가 3개만 존재하지만 'earn' 클래스는 3776개의 문서가 있다. NIPS 자료 집합은 처음 Yann이 NIPS에 제출된 문서의 OCR을 통해 수집하였다. Yann의 자료 집합 역시 불균등하게 분포되어 있다. 이후 tarball이 수집한 문서를 Roweis가 보강하여 균등 집합을 만들었다. Yann의 문서 집합은 총 100여개의 문서로 구성되어 있으며 Roweis의 문서 집합은 1600여개의 문서로 구성되어 있다. NIPS는 학회의 논문을 수집한 집합이기 때문에 Roweis 집합의 경우 Reuters-21578 집합의 약 4.5배에 가까운 54000여개의 많은 단어(불용어 제외)를 가지고 있다. Yann의 집합은 13500여개의 단어를 가지고 있다. 적은 문서에서 많은 단어를 사용하면 클래스를 구분하기 어렵기 때문에 좋은 특징을 선택하는 것은 더욱 어렵다. 그래서 본 논문에서 제안한 방법이 최악의 환경에서 다른 방법에 비해 어느정도 성능 차이가 나는지 비교 분석하였다.

나이브 베이지안[8]을 이용하여 정보 획득(IG), χ^2 방법(X^2)과 본 논문에서 제안한 PWM 방법의 3가지 특징선택 방법을 비교 실험하였다.

5.1 특징 선택 방법 별 성향 분석

각 방법들에 대해 어떤 분포적 특성이 가지고 있는지 파악하기 위해 성향 분석을 하였다. 분석을 하기 위해 Reuters-21578 자료 집합에서 11817개의 문서 중 2000개의 특징을 선택해 특성을 분석하였다.

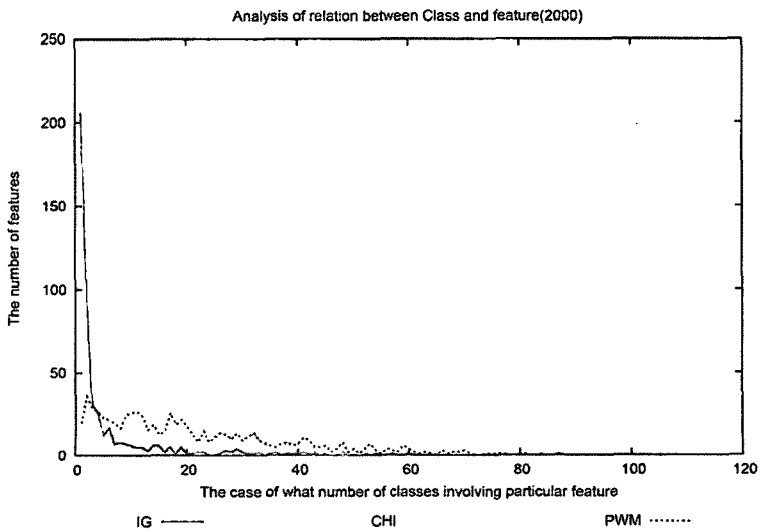


그림 4 Reuters-21578의 클래스와 특징들 사이의 관계 분석

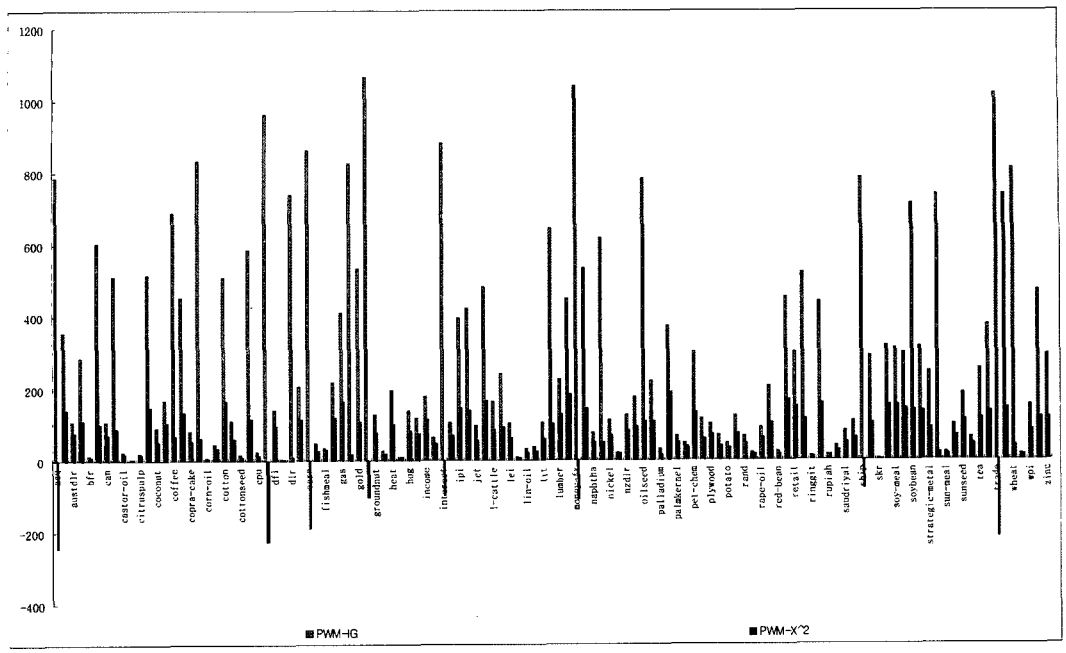


그림 5 Reuters-21578의 클래스별 확보된 상대적 특징 수 비교

첫 번째 실험은 특정 클래스를 대표할 수 있는 특징들이 좋은 특징임을 보이기 위해 각 방법 별 순수도 현황이다. x축은 임의의 특징이 몇 개의 클래스에 나타났는지에 따른 경우를 의미하며, y축은 각 방법들에 의해 선택된 특징들 중에서 x의 경우를 만족하는 특징들의 수를 의미한다. 예를 들어 위 그래프 상에서 정보획득 방법의 경우 오직 하나의 클래스에서만 나타난 특징들의 수는 206개가 된다. 그리고 18개의 클래스에 나타난 특징은 1개만 존재함을 알 수 있다. 이 그래프를 통해 우리가 알 수 있는 것은 정보획득 방법 경우, 적은 클래스에 나타나고 그 클래스를 대표할 수 있는 단어들이 많이 뽑혔다. χ^2 방법은 정보획득 방법과 달리 여러 클래스에 나타난 단어들을 중심으로 특징들을 선택하였다. *PWM* 방법도 χ^2 방법과 비슷한 결과를 얻었지만 그 성향을 많이 틀리다. *PWM*은 정보획득처럼 적은 클래스에 나타난 단어들을 중심으로 선택하고, 선택된 특징의 수는 점점 감소함을 알 수 있다. 그래서 *PWM* 방법은 χ^2보다 클래스를 대표할 수 있는 단어들이 많이 확보되었다. 본 논문에서 제안한 순수도를 고려한 특징 선택으로 가장 좋은 방법은 정보획득 방법이다. 하지만 순수도만 고려했을 경우 클래스마다 대표할 수 있는 특징들의 수는 달라질 수 있다. 만약 균등하게 특징이 선택되지 않고 하나의 클래스에 치우친 형태로 특징들이 선택된다면, 특징들이 많이 확보된 클래스는 좋은 성능을 보이지만 상대적으로 적은 특징을 확보한 클래스는 분류 시 클래스 평가를 할 때 부족한 정보로 올바른 평가를 할 수 없다. 그래서 클래스마다 균등하게 특징들이 선택되어야 한다. 다음 실험에서 클래스마다 대표할 수 있는 특징들이 얼마나 확보되었는지 확인할 수 있다.

두 번째 실험은 선택된 특징들이 클래스별 얼마나 확보되는지 확인 하였다. x축은 클래스를 나타내며 y축은 *PWM*에 대해 상대적인 특징 수를 의미한다. 그래서 y축의 값이 양의 정수이면 *PWM*이 클래스를 나타낼 수 있는 특징을 보다 많이 확보하고 있음을 알 수 있다. 위 실험에서 보는 것과 같이 정보획득 방법인 경우 클래스마다 나타낼 수 있는 특징의 수가 부족함을 알 수 있다. 물론 본 실험에서 여러 클래스에서 나타난 특징은 그래프 상에 모두 반영되었다. 그래서 상대적으로 정보획득 방법이 적은 특징이 선택된 것처럼 보일 수 있다. 하지만, 분류 시 클래스 평가를 위한 특징들이 얼마나 확보되었는지 정량적으로 평가할 수 있다. 위 그래프를 통해 정보획득 방법은 다소 클래스를 나타내는 특징들이 적으며, 특정 클래스는 나타낼 수 있는 특징이 하나도 선택되지 않은 경우도 발생하였다. 하지만 *PWM*은 모든 클래스에 대해 클래스를 대표할 수 있는 단어를 선택하고 있다. χ^2 방법은 *PWM*에 비해 전체적으로 비슷한 성향으로 균등한 특징들을 선별하고 있다. 하지만, 이전의 실험

에서 클래스를 대표할 수 있는 주요 단어들이 많이 선택되지 않았다.

위 실험들을 통해 우리는 정보획득 방법과 χ^2 방법의 성향을 분석 하였으며, *PWM*과 비교 분석하였다. 본 논문에서 좋은 특징이란 어떤 것인지 탐구하였고 성능 비교분석을 통해 그 성능을 입증하였다. 정보획득방법은 클래스를 대표할 수 있는 좋은 특징들이 선별되지만, 선별된 특징들 중에는 몇몇 클래스를 대표할 수 있는 특징들이 부족해 좋은 성능을 내지 못했음을 알 수 있다. 반면에 χ^2 방법은 모든 클래스에 대해 대표할 수 있는 특징들은 선별되지만, 여러 클래스에 나타나는 특징들이 많이 선택되고 분류기의 성능을 극대화시키지는 못하고 있다. 하지만 *PWM*은 클래스를 대표할 수 있는 특징들을 우선적으로 선택하고 클래스마다 균등한 특징 수를 가지도록 하고 있다.

5.2 불균등 분포에서 문서 수에 따른 성능 비교

불균등 분포에서 문서 수를 달리하면서 정확도의 성능이 어떻게 변하는지 실험하였다. 그림 6은 Reuters 자료 집합에서 학습문서의 수를 109, 567, 1191, 2370, 3558, 4707개로 하여 실험한 결과를 그래프이다. 각

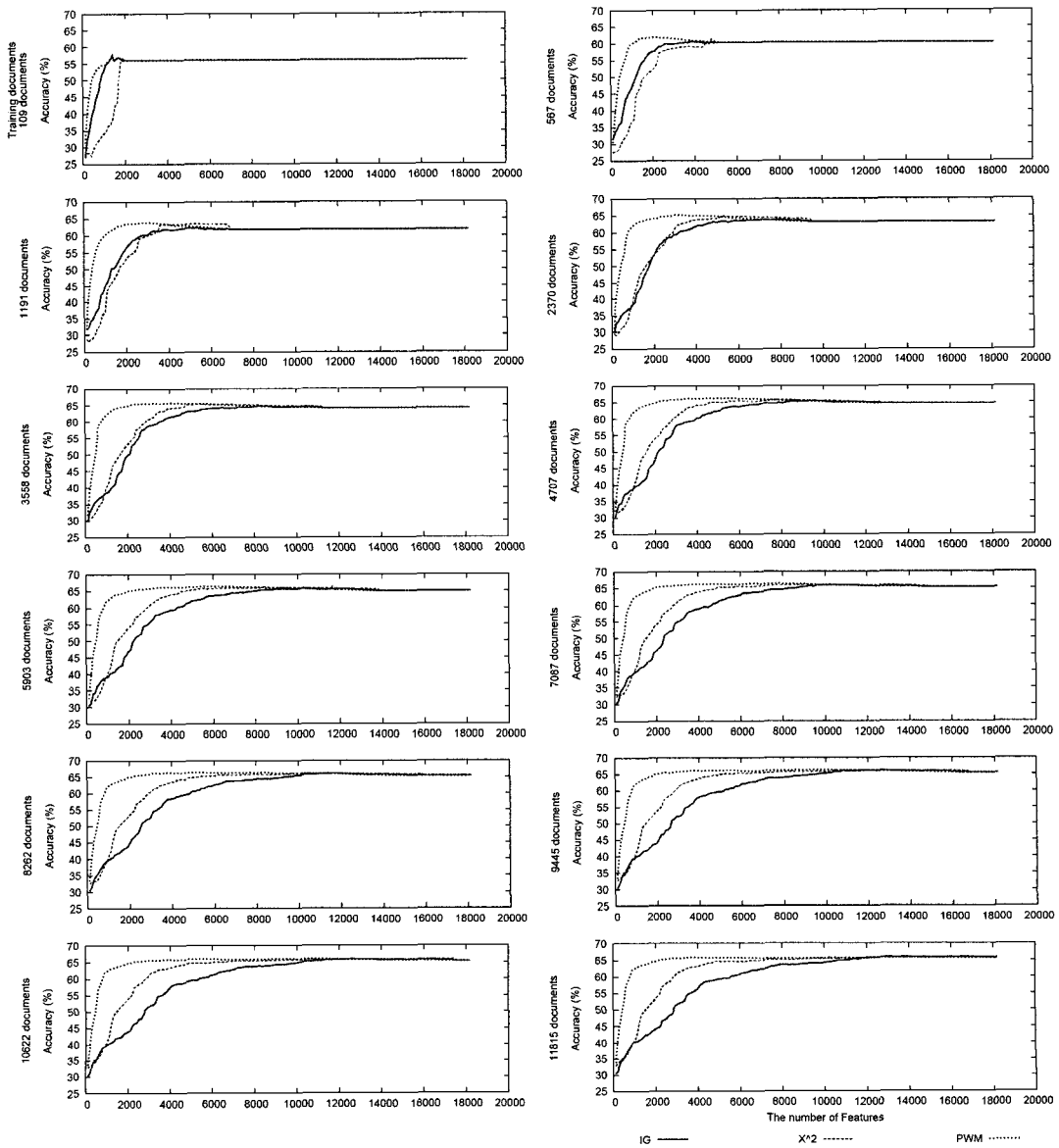


그림 6 Reuters-21578 자료 집합에 대해서 학습 문서 수에 따른 정확도 성능 변화

문서 수에 대해 특징의 수가 증가함에 따라 3가지 방법의 문서분류 정확도 변화를 측정하였다. 각 그래프의 x축은 특징의 수를 나타내고 문서의 수가 증가하면 자연히 특징의 수도 커지게 된다. y축은 문서 분류의 정확도를 나타낸다. 이 그래프에서 109개의 문서를 학습한 경우는 χ^2 방법이 가장 낮은 성능을 보이고 정보 획득 방법이 좋은 성능을 보인다. 하지만 전체적으로 *PWM* 방법이 가장 좋은 성능을 보임을 알 수 있다. *PWM* 방법은 모든 클래스에 필요한 특징을 골고루 확보하고 동시에 불순도를 고려해 클래스를 대표할 수 있는 특징만을 고른다. 다음으로 587개 문서와 1191개 문서에서도 큰 변화 없이 *PWM* 방법이 좋은 성능을 보여준다. 2370개 문서에서는 *PWM* 방법이 독보적으로 좋은 성능을 보이며, 정보 획득 방법과 χ^2 방법이 비슷한 성능을 보이면서 특징 수가 증가함에 따라 χ^2 방법이 정보 획득 방법보다 더 좋은 성능을 보인다. 3558개 문서에서는 χ^2 방법의 성능이 정보 획득 방법의 성능을 앞서가기 시작한다. 반면, *PWM* 방법은 지속적으로 높은 성능을 보임을 알 수 있다.

그림 7은 NIPS-Yann 자료 집합에 대해서 특징의 수가 증가함에 따라 변화는 정확도를 측정한 그래프이다. 전체적으로 Reuters 자료집합과 달리 정보획득 방법이 χ^2방법 보다 좋은 성능을 보이고 있다. 이러한 현상은 자료 집합이 문서의 수는 적으면서 상대적으로 너무 많은 단어들을 내포하고 있기 때문에 다른 클래스에서 나타나지 않는 단어를 선택하는 정보획득 방법이 더 좋은 성능을 보인다고 분석된다. 그림 6에서 적은 문서를 학습했을 때 정보획득 방법이 더 좋았던 것과 같이 NIPS-Yann 자료 집합이 다양한 정

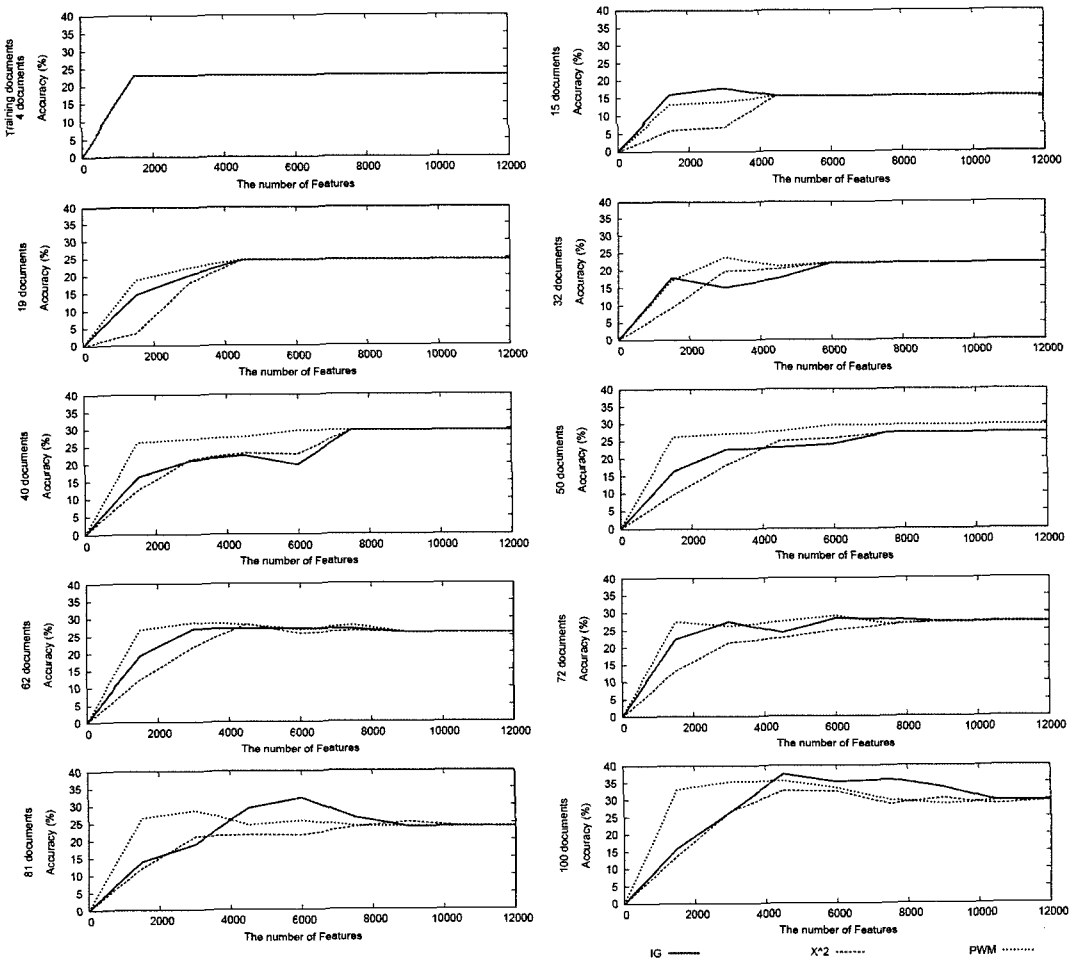
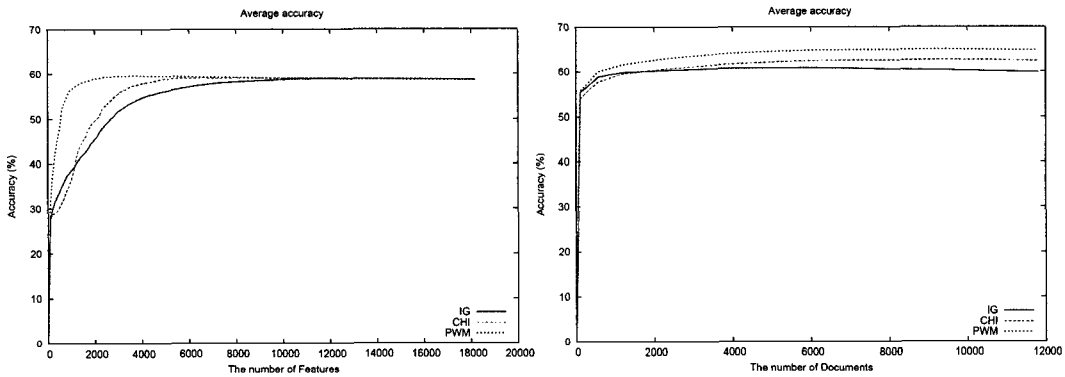


그림 7 NIPS-Yann 자료 집합에 대해서 학습 문서 수에 따른 정확도 성능 변화

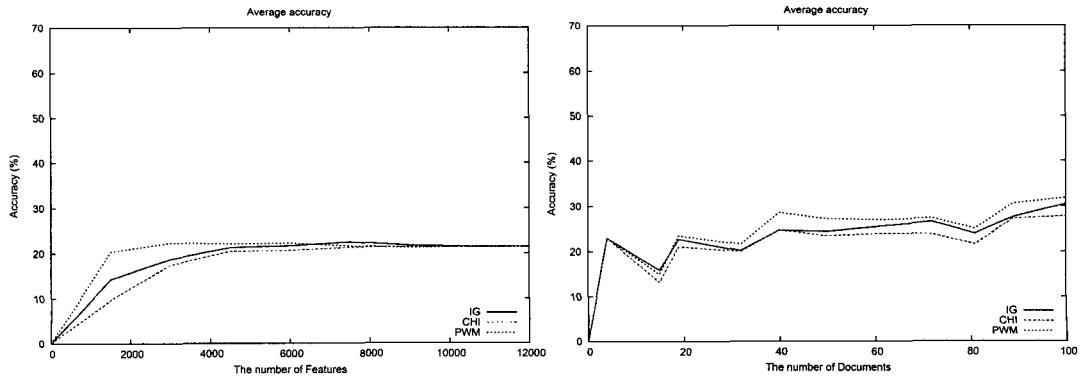
보를 포함하고 있기 때문에 정보획득 방법이 더 좋은 성능을 보였다. 학습 문서가 증가함에 따라 χ^2 방법이 점차 좋은 성능을 보이는 형상만 나타날 뿐이다. 그에 반해 *PWM* 방법은 Reuter 자료집합에서와 같이 적은 특징의 수에서도 좋은 성능을 보이고 있다.

그림 8은 Reuters 자료 집합에 대한 평균 성능을 나타낸 그래프이다. 평균적으로 특징의 수가 증가함에 따라 전체적으로 시스템 향상이 있었다. 하지만 *PWM* 방법이 보다 적은 특징만으로도 좋은 성능을 보임을 알 수 있다. 이와 같은 현상은 *PWM* 방법이 클래스 별로 균등하게 특징을 선택함으로써 분류할 수 있는 기준을 보다 많이 확보하기 때문으로 분석한다. 그 이유는 정보획득방법의 경우 특정 클래스에서만 나타나는 단어를 선택함에 있어서 클래스별로 균등하게 선택하지 않는다. 그래서 총 100개의 특징을 선별한다고 할 때 최악의 경우 선택된 100개의 특징이 특정 하나의 클래스에서만 나타난 단어일 수도 있다. 이러한 단어는 하나의 클래스를 구분하기에는 너무 많은 정보를 내포하게 되지만 반대로 다른 클래스에 대해 분류기준이 될 수 있는 특징들이 존재하지 않을 수 있다. 오히려 10개의 특징만으로 클래스를 나타낼 수 있음에도 불구하고 불필요한 정보가 특징으로 선택될 수 있다. 그래서 문서 수가 증가함은 여러 클래스에 균등하게 증가하는 것이 아니라 특정 클래스에 국한되어 증가함으로써 정보획득 방법이 점점 안좋은 성능을 보임을 알 수 있다. χ^2 방법은 정보획득 방법과 달리 문서의 단어들이 어떤 의존 관계를 가지고 있는지를 측정함으로써 주요 특징인지를 구분한다. 이와 같은 정보는 적은 자료에서는 판단하기 힘들며 보다 많은 정보를 요구하게 된다. 그래서 적은 문서의 집합에서는 매우 다양한 단어들이 나타나 성향을 파악하기 힘들다. 하지만 자료집합에서 사용하고 있는 단어의 종류는 제한적이기 때문에 문서의 수가 증가함에 따라 좋은 성능을 보임을 알 수 있다. *PWM* 방법은 클래스별로 균등하게 특징을 선택하게 함으로써 다른 방법들이 가지고 있는 문제점을 해결하였다.



(a) 특징 수에 대한 성능 (b) 문서 수에 대한 성능

그림 8 Reuters-21578 자료 집합의 평균 정확도



(a) 특징수에 대한 성능 (b) 문서수에 대한 성능

그림 9 NIPS-Yann 자료 집합의 평균 정확도

NIPS-Yann 자료 집합의 성능에서도 *PWM* 방법이 좋은 성능을 보였다. Reuters 자료집합과 다른 점은 χ^2 방법이 정보획득방법보다 낮은 성능을 보인다는 것이다. 이 현상은 Reuters 자료 집합에서와 같이 분포에 따른 의존도를 측정하기 위한 정보가 부족하기 때문으로 분석된다. NIPS-Yann 자료집합은 전체 문서수는 100개이며 보유하고 있는 단어의 수는 무려 12000여개가 된다. 이와 같이 적은 문서에서 다양한 단어 정보를 보유하고 있기 때문에 의존도를 측정한다는 것은 매우 어렵다. 좀 더 자료를 보강한 균등 자료집합 NIPS-Roweis에서 시험 결과를 보면 알 수 있듯이 학습 문서 수가 증가함에 따라 χ^2 방법이 정보획득방법보다 더 좋은 성능을 보임을 알 수 있다. *PWM* 방법은 다양한 단어를 사용하면서 정보가 부족한 열악한 환경에서도 다른 방법들에 비해 더욱 좋은 성능을 보임을 알 수 있다.

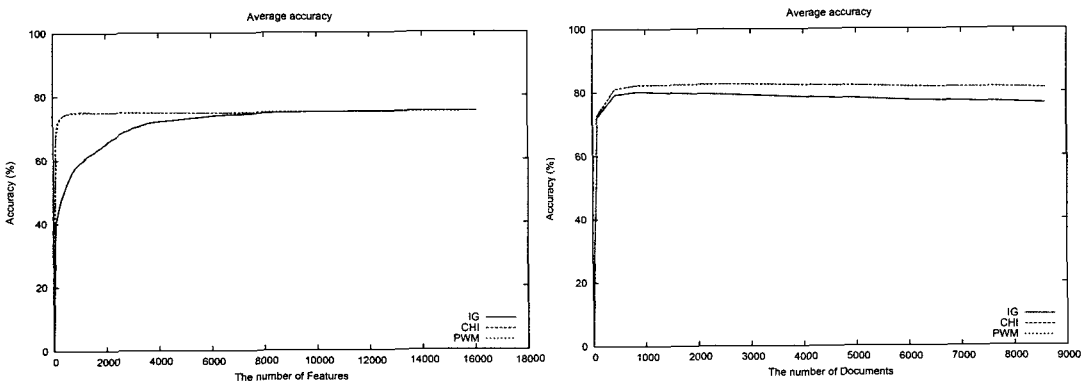
결과적으로 적은 문서를 학습할 경우, 평가하기 어려운 다양한 단어들이 나타남에 따라 클래스를 구분하기 위한 특징을 선택함에 있어서 정보획득 방법은 좋은 성능을 보인다. 하지만 클래스들 마다 속해 있는 정보양이 서로 다른 상태에서 단어들의 의존도 성향을 파악하는 것은 매우 어렵기 때문에 χ^2 방법이 다른 방법들보다 낮은 성능을 보이게 된다. 하지만 문서수의 증가로 기존 단어들의 의존도 성향을 파악가능할 때는 χ^2 방법이 정보획득 방법보다 좋은 성능을 보였다. 더불어 정보획득 방법은 특징 클래스에 치우쳐 특징을 선별할 수 있기 때문에 문서 수의 증가가 성능을 저하 시킬 수도 있다. 이와 같은 문제점들을 해결한 것이 *PWM* 방법이다. *PWM* 방법은 순수도에 따라 클래스를 대표할 수 있는 특징을 선별하며 그와 동시에 클래스별로 균등하게 선택함으로써 시스템 전체의 성능을 높일 수 있다.

본 논문에서 제안한 좋은 특징이란 (1) "하나의 클래스에만 나타난 단어, 즉 순수도가 높은 단어가 좋은 단어이다." 와 (2) "클래스마다 균등하게 특징이 선택되어야 한다." 라는 것이 실험을 통해 입증 되었다.

5.3 균등 분포에서 상대적 성능 평가

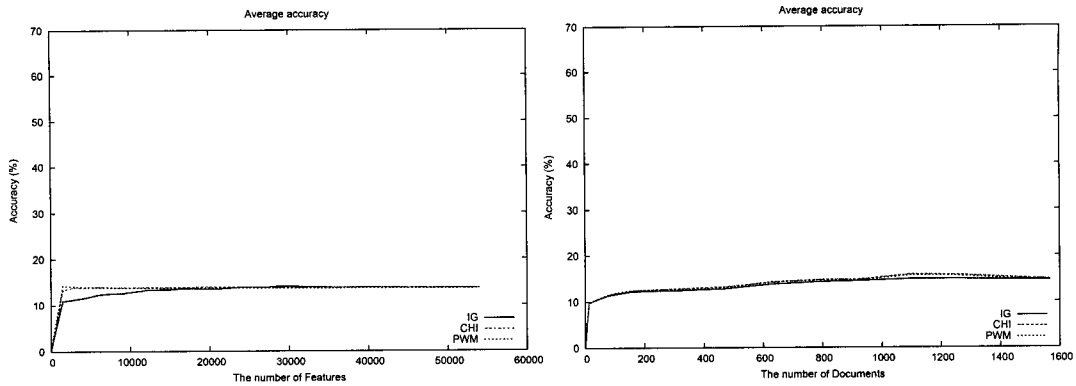
Reuters-21578 자료 집합은 불균등한 자료 분포를 가지고 있다. 그래서 일반적으로 이러한 제약사항을 없애기 위해 10개의 클래스만 선택하여 실험을 하는 경우가 많다. 즉, "trade" ,"earn", "interest", "crude", "ship", "acq", "money-fx", "grain", "corn", "wheat" 클래스만을 이용함으로써 문서 수와 문서의 길이가 균형 있게 분포된 자료 집합으로 실험을 할 수 있다.

그림 10은 각 방법들에 대해 특징 수 변화에 따른 실험과 문서 수 변화에 따른 실험이다. 정보획득방법이 다른 방법들에 대해 낮은 성능을 보였다. 이 실험을 통해 정보획득 방법은 다른 클래스에 널리 퍼져 있지 않는 단어만 고려하는 것은 위험한 요소를 담고 있다는 것을 알 수 있다. 그 요소는 전문용어 또는 잘 사용되지 않는 단어, 오기(誤記)와 같은 단어들이 다른 클래스에서 널리 사용되지 않기 때문에 특징으로 선별될 가능성이 있다. 이러한 단어들은 같은 클래스라 할지라도 여러 문서에 나타나지 않는다. 결과적으로 이런 단어들을 특징을 선별하면 클래스를 판단하기 어렵다. 따라서 단어들의 의존도를 기반으로 한 χ^2 방법이 더욱 좋은 성능을 보인다. *PWM* 방법은 클래스 별로 균등하게 특징을 수집하도록 $\Pr(w|c)$ 가중치를 부여하고 있다. $\Pr(w|c)$은 클래스 c에서 단어 w가 어느 정도 나타나는지에 대한 확률로서 클래스 내부에서의 중요도를 의미한다. 더불어 단순히 클래스 별로 균등하게 특징을 수집하는 것뿐만 아니라 클래스 안에서 널리 사용되는 의미 있는 단어를 특징으로 선별하게 된다. 그래서 정보획득과 같은 문제점을 해결



(a) 특징수에 대한 성능 (b) 문서수에 대한 성능

그림 10 Reuters-21578 자료 집합의 평균 정확도



(a) 특징수에 대한 성능 (b) 문서수에 대한 성능

그림 11 NIPS-Roweis 자료 집합의 평균 정확도

할 수 있다. 그래프를 통해 균등 분포 자료집합에서는 χ^2와 *PWM* 방법의 성능에 큰 차이가 없음을 알 수 있다.

그림 11은 NIPS-Yann 자료집합에 문서를 더 보충하여 균등집합으로 만든 NIPS-Roweis 자료집합에 대한 실험결과이다. 특징 수 변화에 대한 실험결과에서는 정보획득 방법이 조금 성능이 좋지 못하였으며 χ^2 방법과 *PWM* 방법의 성능의 차이가 거의 없었다. 마찬가지로 문서수의에 대한 성능에서 비슷한 성능을 보였다. 다만, 보다 적은 문서, 보다 적은 특징만을 고려했을 때는 불균등 분포 실험에서와 같이 *PWM* 방법이 조금 더 좋은 성능을 보였다.

6. 결론 및 향후 연구과제

본 논문에서는 보다 효율적으로 특징을 선택하기 위해 클래스 별 단어의 불순도와 문서의 불균등 분포를 고려한 특징 선택 방법을 제안하였다. 본 논문에서 제안한 *PWM* 방법은 클래스를 대표할 수 있는 특징 후보들을 단어의 불순도 측정을 통해 얻고, 문서의 불균등 분포를 고려하여 특징을 선택하였고 이를 수식화 하였다. *PWM* 방법은 단어의 불순도를 고려하여 클래스를 대표할 수 있는 단어를 선택한다. 더불어 문서 수, 문서의 길이가 불균등하게 분포되어 있는 자료집합을 고려한 가중치를 부여함으로써 성능을 더욱 높였다. 결과적으로 *PWM* 방법은 다른 클래스에서 널리 사용되지 않는 용어이면서 본 클래스에서는 많이 사용되는 주요 단어를 특징으로 선별할 수 있다.

불균등 분포 자료집합과 균등 분포 자료집합의 실험을 통해 제안한 특징 선택방법이 우수함을 입증하였다. 향후 불순도 및 불균등 분포 자료 집합을 고려한 특징 선택 방법을 텍스트 자료 뿐 아니라 음성 및 영상 자료의 특징 선택에 사용함으로써 불필요한 자료를 제거하는데 이용할 수 있을 것이다. 예를 들어 음성 데이타 중 노이즈 채널 및 음성 채널 필터링을 통해 실제 데이타를 표현할 수 있는 특징으로 선택할 수 있을 것이다. 이러한 다양한 응용 분야에 적용함으로써 학습 효율을 높이고 문서 분류의 정확도를 향상시키는데 많은 기여를 할 것으로 기대된다.

참 고 문 헌

[1] K. Kira and L. A. Rendell. "The feature selection problem: Traditional methods and a new algorithm," In 10th National Conference on Artificial Intelligence, pp. 129-134. MIT Press 1992.

[2] M. Dash, K. Choi, P. Scheuermann, H. Li, "Feature selection for clustering - a filter solution," Proc. of IEEE-ICDM, pp. 115-122, 2002.

[3] R. Caruana, D. Freitag, "Greedy attribute selection," Proc. of ICML94, pp. 28-36, 1994.

[4] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Ph. D diss. Hamilton, NZ: Wailkato University, Department of Computer Science, 1999.

[5] I. H. Witten, E. Frank, Data Mining, Morgan Kaufmann Publishers, 2000.

[6] G. H. John, R. Kohavi, K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proc. of ICML94, pp. 121-129, Morgan Kaufmann Pulishers, San Francisco, CA, 1994.

[7] Y. Yang, J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. of ICML97, pp. 412-420, 1997.

[8] Tom Mitchell, Machine Learning, McGraw Hill, 1996.

[9] D. D. Lewis, "Reuters-21578 Text Categorization Test Collection Distribution 1.0 README file (v 1.3)," 2004, http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt

[10] S. Roweis, "NIPS Conference Papers Vols0-12"

http://www.cs.toronto.edu/~roweis/data/nips12raw_str602.tgz

강 진 범
정보과학회논문지 : 소프트웨어 및 응용
제 34 권 제 6 호 참조

양 재 영
1998년 한양대학교 전자계산학과 졸업(학사). 2000년 한양대학교 대학원 전자계산학과 졸업(석사). 2003년 한양대학교 대학원 컴퓨터공학과 졸업(박사). 2003년~2005년 Openbase 책임연구원. 2005년~2006년 (주)동부정보기술 컨설팅사업부 과장. 2007년~현재 코리아와이즈넛 연구소 솔루션센터 센터장. 관심분야는 지능형 에이전트, 인공지능, 기계학습, 정보검색/추출

최 중 민
정보과학회논문지 : 소프트웨어 및 응용
제 34 권 제 6 호 참조