효율적인 문서 분류를 위한 혼합 특징 집합과 하이브리드 특징 선택 기법

Combined Feature Set and Hybrid Feature Selection Method for Effective Document Classification

인 주 호¹ 김 정 호^{1*} 채 수 환²
Joo-ho In Jung-ho Kim Soo-hoan Chae

요 약

본 연구에서는 효율적인 온 라인 문서 자동 분류를 위해 매우 중요한 분류 작업의 전처리 단계인 특징선택을 위한 새로운 방법이 제안된다. 대부분의 기존 특징선택 방법 연구에서는 특징 집합의 모집단이 단일 모집단으로써 한 모집단이 가지는 정보만으로 분류에 적합한 특징들을 선택하여 특징 집합을 구성하였다. 본 연구에서는 단일 모집단에 한하여 수행되는 특징선택 뿐 만 아니라, 다중 모집단을 가지는 혼합 특징 집합에 대해서 특징선택을 함으로써 다양한 정보를 바탕으로 한 특징 집합을 구성하였다. 혼합 특징 집합은 두 종류의 특징 집합으로 구성된다. 즉 각각 문서로부터 추출한 단어로 구성된 원본 특징 집합과 원본 특징 집합으로부터 LSA를 이용하여 새로 생성한 변형 특징 집합이다. 혼합 특징 집합으로부터 필터 방법과 래퍼 방법을 이용한 하이브리드 방식의 특징 선택을 통해 최적의 특징 집합을 찾고, 이를 이용하여 문서 분류 실험을 수행하였다. 다양한 모집단의 특징들의 정보를 모두 고려함으로써 보다 향상된 분류 성능을 보일 것이라고 기대하였고, 인터넷 뉴스 기사를 대상으로 분류 실험한 결과 90% 이상의 향상된 분류 성능을 확인하였다. 특히, 재현율과 정밀도 모두 90%이상의 성능을 보였으며, 둘 사이의 편차가 낮은 것을 확인하였다.

주제어 : 문서 분류;특징 선택;혼합 특징 집합;LSA;하이브리드 특징 선택

ABSTRACT

A novel approach for the feature selection is proposed, which is the important preprocessing task of on-line document classification. In previous researches, the features based on information from their single population for feature selection task have been selected. In this paper, a mixed feature set is constructed by selecting features from multi-population as well as single population based on various information. The mixed feature set consists of two feature sets: the original feature set that is made up of words on documents and the transformed feature set that is made up of features generated by LSA. The hybrid feature selection method using both filter and wrapper method is used to obtain optimal features set from the mixed feature set. We performed classification experiments using the obtained optimal feature sets. As a result of the experiments, our expectation that our approach makes better performance of classification is verified, which is over 90% accuracy. In particular, it is confirmed that our approach has over 90% recall and precision that have a low deviation between categories.

keyword : document classification;feature selection;mixed feature set;LSA;hybrid feature selection;

1. 서 론

본 연구에서는 분류에 최적화된 특징 집합을 찾는 특징 선택(feature selection)에 초점을 맞춘다. 특징 선택은

문서 분류 뿐 만 아니라, 다양한 분류 분야 및 인식 분야에서 핵심이 되는 작업이다. 문서를 가장 잘 표현하면서도 서로 다른 범주를 잘 구별시켜 주는 최적의 특징 집합을 찾는 것이다. 분류에 최적화된 특징 집합은 분류 성능을 높이며 분류 속도도 향상시킨다. 즉, 특징 선택은 특징 집합의 크기를 줄임으로써 차원의 저주(curse of dimension)를 해결하고 분별력 있는 특징들로 구성함으로써 분류 성능을 보장한다. 대부분의 특징 선택 연구는 이러한 목적을 달성하기 위해 다양한 방법들을 제안하였다(Guyon & Elisseeff, 2003; Dasgupta et al., 2007).

그러나 기존 특징 선택 연구들은 문서로부터 추출한

¹ Computer Engineering, Korea Aerospace Univ., Gyeonggi-do, 412-791, Korea.

² School of Electronics Telecommunication & Computer Engineering, Korea Aerospace Univ., Gyeonggi-do, 412-791, Korea.

* Corresponding author(natul2@kau.ac.kr)

[Received 11 December 2012, Reviewed 27 December 2012(R2 24 April 2013, Accepted 28 June 2013)]

특징들을 정제하지 않고 사용한다. 문서를 표현하고자 하는 특징은 보통 문서 내에 포함된 단어이다. 단어를 특징으로 문서들을 표현한 단어-문서 행렬은 매우 희소한 행렬(sparse matrix)이다. 단어-문서 행렬이 희소행렬일 경우 제대로 된 정보를 얻어 올 수 없기 때문에 앞선 특징 선택 기법들을 사용하여 얻은 값을 신뢰할 수 없다.

희소 행렬 문제를 해결하면서 동시에 단어의 의미를 고려할 수 있는 방법으로 LSA(Latent Semantic Analysis)가 있다. LSA는 단어-문서 행렬로부터 새로운 특징 집합을 생성한다. 생성된 특징 집합은 기존 특징 집합보다 축소된 크기를 가지며 새로운 특징 집합으로 구성된 문서 행렬은 밀집 행렬(dense matrix)이 된다. 또한, 기존 특징인 단어 간의 공기 정보를 바탕으로 단어의 의미를 새로운 특징 집합에 반영한다. LSA의 단어 간 공기(concurrence) 정보를 통한 의미 반영은 동의어 연구나 군집화(clustering) 연구에서 입증되었다(Landauer & Dumais, 1997; Deerwester et al., 1990). 그러나 기본 LSA는 범주가 사전에 정의된 분류에는 적합하지 못하다. 이러한 문제를 해결하기 위해 지도적 학습을 통한 지도적 LSA가 제안되었다(Chakraborti et al., 2006; Sun et al., 2004).

특징 집합의 모집단의 종류의 변화를 통한 새로운 특징 선택에도 불구하고 여전히 한계가 존재한다. 특징 선택의 대상이 되는 특징 집합은 오직 단일 모집단을 가지기 때문이다. 다양한 모집단의 특징 집합으로부터 얻을 수 있는 정보들을 고려하지 못하고 편향된 정보만으로 특징을 선택하면 그 만큼의 정보를 손실하게 된다.

본 논문은 분류에 적합하며 다양한 모집단의 특징 집합의 정보를 고려한 특징 선택을 위해 혼합 특징 집합을 구성하고, 이로부터 최적의 특징들을 선택하기 위한 특징 선택 방법인 필터(filter)와 래퍼(wrapper) 방법을 혼합한 하이브리드 특징 선택 방법을 제안한다. 혼합 특징 집합은 분류 대상인 문서 내의 단어로 구성된 원본 특징 집합과 지도적 LSA 기법을 통해 구한 새로운 변형 특징 집합을 혼합하여 구성하였다. 혼합 특징 집합의 각 집합으로부터 필터 방법을 이용하여 분별력 있는 특징들만 선택하여 새로운 특징 집합을 만들고 이를 혼합한 하나의 혼합 특징 집합으로부터 래퍼 방법을 이용하여 최적의 특징 집합을 선택하였다. 이때 사용한 래퍼 방법으로 유전 알고리즘(GA: Genetic Algorithm)을 사용하였다. 유전 알고리즘은 최적의 해를 탐색하는데 매우 좋은 성능을 보이며 최적의 해에 빠르게 수렴하는 장점이 있으므로 최적의 특징 집합을 탐색하는 문제에 매우 적합하다. 이를 특징 선택 기법에 적합하도록 변형하여 적용하였다. 제안하

는 방법을 인터넷 뉴스 문서에 적용하여 실험하였다.

2. 관련 연구

특징 선택은 분류 작업 시 분류의 대상을 표현하는 전체 특징들 중 그 대상을 가장 잘 표현하는 특징들만을 선택하여 부분 특징 집합을 구성하는 역할을 하며, 현재 다양한 분류 연구에서 연구되어지고 있는 세부 연구 주제이다. 분류 성능을 악화시키는 중복된 정보를 가지는 특징과 불필요한 특징을 전체 특징 집합으로부터 제거하여 크기가 작은 최적의 특징 집합을 구성한다. 이 집합을 이용하여 분류함으로써 기존 전체 특징 집합을 이용하였을 때 보다 빠르고 더 정확한 분류 성능을 얻는 것이 특징 선택의 목표이다. 분류하고자 하는 데이터의 양이 기하급수적으로 증가하는 요즈음의 추세로 볼 때, 효과적인 특징 선택은 더욱 중요해진다.

(그림 1) 특징 선택의 절차

(Figure 1) Procedures of the Feature Selection

최적의 특징 집합을 구성하기 위한 특징 선택은 세 단계로 구성된다. 각각 전체 특징 집합으로부터 1) 후보 특징 집합을 생성(generation)하는 단계, 2) 후보 특징 집합을 평가(evaluation)하는 단계, 마지막으로 3) 최종적으로 선택된 부분 특징 집합에 대한 검증(validation) 단계이다. 위 세 단계에 따른 특징 선택 과정은 그림 1에 나타났다.

첫 번째 단계인 후보 특징 집합 생성을 위한 기법들은 크게 완전탐색(complete search), 순차탐색(sequential search), 및 임의탐색(random search)이 있다(Liu & Yu, 2005). 완전탐색은 원본 특징 집합의 크기가 클수록 생성할 수 있는 후보 특징 집합의 수가 기하급수적이라 대부분 순차탐색 또는 임의탐색 방식으로 후보 특징 집합을 생성한다.

생성한 후보 특징 집합이 분류에 적합한지 여부를 확인하기 위해 평가를 한다. 평가 방법과 시기에 따라 총 세 가지 특징 선택 방법이 존재한다. 특징 집합의 개개별 특

징의 독립된 정보만으로 평가하는 필터 방법, 특징 분류기의 분류 성능을 평가 척도로 사용하는 래퍼 방법, 그리고 기존 기계 학습 시 동시에 특징 선택이 이루어지는 임베디드(Embedded) 방법이 있다.

필터 방법은 문서가 가지고 있는 정보를 바탕으로 후보 특징 집합을 평가한다. 특징 집합으로 표현된 문서 벡터들 간의 거리나 IG(Information Gain)과 같은 측정치를 계산하여 이를 기준으로 특징의 선택 여부를 결정한다. 일반적인 측정치로는 DF(Document Frequency), MI(Mutual Information), Chi, TS(Term strength), Fuzzy rank 등이 있다(Chen et al., 2009; Selamat & Omatu, 2004; Yang & Pedersen, 1997; Peng et al., 2005). 필터 방법은 독립된 단일 특징들의 가중치만을 가지고 특징을 선택하기 때문에 특징들간의 관계를 고려하지 못하는 한계가 있다.

래퍼 방법은 문서 분류에 사용할 분류기의 특성을 고려하여 분류기에 적합한 특징을 선택하기 위해 학습 데이터를 분류기로 분류하여 얻은 분류 성능으로 특징 집합을 평가한다. 래퍼 방법은 모든 후보 특징 집합들에 대한 분류를 수행하기 때문에 필터 방법에 비해 계산 복잡도가 크다. 하지만 분류기의 특성이 고려되기 때문에 필터 방법에 비해 분류 정확도가 높다(John et al., 1994).

임베디드 방법은 래퍼 방법과 동일하게 특징 집합을 평가하지만 평가 시기가 분류기 학습과 동시에 일어난다. 분류기의 학습과 동시에 특징 선택이 수행되기 때문에 래퍼 방법의 큰 시간 복잡도를 다소 해결할 수 있지만 여전히 분류기를 통한 분류 정확도로 특징 집합을 평가하기 때문에 비싼 작업이다.

이러한 이유로 필터 방법과 래퍼 방법을 혼합한 하이브리드 특징 선택 방법에 대한 연구가 진행되었다(Luukka, 2011; Gheyas & Smith, 2010). 이는 필터 방법의 간편함과 래퍼 방법의 정확도를 통해 기존 단일 방법 보다 효율적인 특징 선택을 목표로 한다. 하지만 독립된 단일 특징의 특성만을 고려하여 단어 간의 관계를 의미 분석의 부재는 여전히 존재한다. 이를 위해 본 연구의 선행 연구에서 지도적 LSA를 이용하여 단어 간의 관계를 바탕으로 새로운 특징 집합을 생성하고, 생성한 특징 집합을 기준으로 래퍼 방법으로 최적의 특징 집합을 생성하였다(Jung-ho et al., 2012). 단어 간의 관계를 통해 생성된 새로운 특징 집합은 기존 특징 집합이 가지고 있는 잠재된 의미를 반영하기 때문에 특징이 가지는 중의성을 해소하여 향상된 분류 결과를 보였다.

앞의 모든 관련 연구는 특징 집합의 모집단이 단일 모집단으로 특징이 가지는 다양한 정보를 모두 고려하지

못하였다. 원본 특징 집합에 한에서만 특징 선택이 수행되었으며(Liu et al., 2011), 선행 연구에서도 LSA를 통해 생성한 새로운 특징 집합만으로 특징 선택을 수행하였기 때문에 특징 집합은 단일 모집단을 가진다. 본 연구는 단일 모집단이 아닌 다중 모집단을 가지는 혼합 특징 집합에 대한 특징 선택을 수행하여 단일 모집단이 가지지 못하는 다양한 특징 정보를 모두 포함하여 특징 선택을 수행하였다.

3. 제안하는 연구방법

3.1 혼합 특징 집합

본 연구의 핵심은 다중 모집단을 가지는 혼합 특징 집합으로부터 최적의 특징 집합을 선택하는 것이다. 혼합할 특징 집합은 1) 원본 특징 집합과 2) 변형 특징 집합이다. 원본 특징 집합은 문서로부터 추출한 단어들로 구성된 집합이며, 이는 문서에서 표현적으로 보이는 정보를 담고 있다. 변형 특징 집합은 LSA를 통해 생성된 새로운 특징 집합이며, 이는 단어의 공기 정보를 바탕으로 단어 간의 관계 정보를 담은 새로운 특징 집합이다. 즉, 원본 특징 집합이 가지는 형태론적, 표면적 정보와 변형 특징 집합이 가지는 잠재된 의미 정보를 모두 고려하기 위해 두 특징 집합을 혼합하여 특징 선택을 수행한다.

원본 특징 집합의 요소는 문서 내에 존재하는 단어이며, 본 연구에서는 체언 중 수사를 제외한 명사, 대명사, 그리고 용언 중 형용사와 동사 품사를 가지는 단어이다. 이는 문서에서 표현하고자 하는 정보를 담기 위해 사용되는 주된 품사이므로 위 품사를 가지는 단어를 요소로 원본 특징 집합을 구성한다.

변형 특징 집합은 LSA를 이용하여 문서 내의 단어 간의 공기 정보를 바탕으로 새로운 특징 집합을 생성한다. LSA는 행렬 분해 기법인 SVD(Singular Value Decomposition)를 통해 원본 특징 집합으로부터 새로운 특징 집합을 생성하며, 이 때 불필요한 특징들을 제거함으로써 특징 집합의 크기를 줄인다. SVD는 식 (1)과 같이 단어-문서 행렬(Amxn)을 세 개의 행렬로 분해하며, 각각 단어-단어의 상관관계를 나타내는 행렬(Umxn), 문서-문서 상관관계를 나타내는 행렬(Vnvn), 그리고 두 행렬의 특이치(singular value)를 가지는 대각 행렬(Snvn)이다.

$$A = USV^T \quad (1)$$

즉, 행렬 U 는 단어 간의 상관관계를 내포하는 새로운 의미 공간이고, 행렬 V 는 문서 간의 상관관계를 내포하는 새로운 의미 공간이다. 그리고 행렬 S 의 특이치는 행렬 U 와 V 의 기저 벡터가 의미 공간에서 가지는 강도를 의미하며, SVD의 차원 축소는 이 특이치를 기준으로 이루어진다. 특이치가 낮은 기저 벡터는 불필요한 차원으로 간주하여 제거하고 상위 k 개의 특이치에 대응하는 열 벡터만 남긴 행렬 U_k 와 V_k 를 생성한다.

특징 추출에서는 SVD의 행렬 분해 결과 중 단어-단어 상관관계를 나타내는 행렬 U 에 초점을 둔다. 기존 단어-문서 행렬을 행렬 U 에 투영함으로써 단어의 의미를 반영한 문서 벡터를 구할 수 있다. 식 (2)로 단어-문서 행렬을 행렬 U 로 투영하여 새로운 특징-문서 행렬 C 를 생성한다.

$$C = AU_k \quad (2)$$

차원 축소를 통해 잡음을 제거하여 문서 분류의 계산량(computation cost)을 줄이고 좀 더 정확한 의미 분석을 가능하게 한다. 이러한 장점 때문에 LSA는 정보 검색의 키워드 인덱싱 및 클러스터링 등의 다양한 연구 분야에서 사용된다.

하지만 LSA는 문서 분류에는 적합하지 않다. 문서 분류는 사전에 정해진 범주로 문서를 분류하는 것으로, 특징 추출 시 서로 다른 범주간의 높은 차별성과 범주간의 높은 결합성을 고려해야 한다. LSA는 특징 추출 시 오직 특이치만 고려하고 위 두 가지 사항을 제대로 고려하지 않기 때문에 높은 분류 성능을 기대하기 어렵다. 이러한 이유로 몇몇 분류 연구에서 분류에 적합한 특징 추출을 위해 지도적 학습을 추가한 다양한 지도적 LSA 방법을 제안하였다.

그 중 Chakraborti(2006)는 기존 단어-문서 행렬에 범주 정보를 추가로 삽입하는 방법으로 SVD 행렬 분해 시 범주 정보 역시 포함 되도록 하는 Sprinkling 방법을 제안하였다. 그리고 Sun(2004)은 단어-단어 행렬인 U 에서 서로 다른 범주간의 거리 차이를 기준으로 차원을 선택하는 방법을 제안하였다. 본 논문 역시 분류를 위해 지도적 학습이 추가된 LSA를 사용하며, 위 두 가지 방법 중 Sprinkling 을 사용하였다.

3.2 혼합 특징 집합에 대한 특징 선택 방법

본 연구에서 적용한 특징 선택 방법과 그에 따른 처리 과정은 (그림 2)와 같다.

(그림 2) 혼합 특징 집합에 대한 특징 선택 절차
(Figure 2) Procedures of the Feature Selection for the Hybrid Feature Set

두 특징 집합을 혼합하기에 앞서 먼저 각각에 대해 필터 방법을 적용하여 1 차 특징 선택 작업을 수행한다. 이는 각 특징 집합 내에서도 분류에 적합한 상위 특징들만 빠르게 추려내어 최종 특징 집합 선택 시의 속도 개선을 위함이다. 이때 사용하는 필터 방법은 F값(F-measure)으로, 이는 독립된 두 집단 간의 분산크기 대 집단 내 분산크기의 비(ratio)이다. 즉, F값 이 크면 클수록 비교집단들 사이에 차이가 존재할 가능성이 높다고 판단하므로, 한 특징상의 서로 다른 범주의 문서들의 F 값을 구하여 F 값 이 큰 해당 특징이 분류를 위한 분별력이 높은 특징으로 판단하여 선택한다. 전체 특징 집합 중 하나의 특징에 대한 두 범주의 문서 집합 X 이 있을 때, 다음 F값을 구하는 식 (3)을 통해 F값이 일정 경계 값 이상인 특징을 선택하여 특징 집합을 구성한다.

$$F = \frac{MS_{between}}{MS_{within}} \quad (3)$$

$$MS_{between} = \sum_{j=1}^k (\bar{X}_j - \bar{X})^2 / (k-1)$$

$$MS_{within} = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 / (N-k)$$

j : 집단 수 (=2)

$\bar{X}$: 전체 평균

$\bar{X}_j$: j 번째 집단의 평균

X_{ij} : j 번째 집단의 i 번째 문서의 값

N : 전체 집단의 표본 수

n_j : j 번째 집단의 표본 수

k : 집단의 수 (= 2)

필터 방법을 원본 특징 집합 A 와 변형 특징 집합 B 에 적용한 결과인 두 부분 특징 집합 A' 과 B' 을 혼합하여 하나의 특징 집합 $F(=A'UB')$ 을 구성한다. 특징 집합 F 는 서로 다른 두 모집단의 특징이 결합된 특징 집합으로 원본 특징 집합이 가지는 정보와 변형된 특징 집합이 가

는 정보를 모두 포함한다. 이를 최적화 알고리즘인 유전 알고리즘을 적용한 래퍼 방법으로 최종 특징 집합을 선택한다. 본 연구에서 래퍼 방법의 분류기로는 SVM (Support Vector Machine)을 사용한다. (그림 3)은 본 연구의 래퍼 방법의 특징 선택 작업 처리 과정을 보인다.

(그림 3) 유전 알고리즘을 이용한 특징선택 절차
(Figure 3) Procedures of the Feature Selection using the Genetic Algorithm

4. 문서 분류 실험 및 결과

4.1 실험 개요

전체 분류 실험을 위한 실험 시스템 구성은 (그림 4)와 같이 구성하였다. 본 연구에서 제안한 특징 선택 방법을 통해 구해진 최종 특징 집합으로 실험 데이터를 분류함으로써 그 효과를 검증하였다.

(그림 4) 문서 분류 시스템 구조
(Figure 4) Structure of the Document Classification System

본 연구에서는 문서 분류 실험을 위해 조선 일보의 인터넷 기사를 사용하였다. ‘정치’ 기사와 ‘경제’ 기사로

분류 할 두 범주를 정하여, (표 1)에서 보이는 바와 같이 학습을 위한 학습 데이터와 실제 분류 성능을 검증하기 위한 실험 데이터를 각각 구성하였다. ‘정치’와 ‘경제’ 두 범주의 기사에는 유사하거나 중복되는 단어가 많이 존재하므로, 특징 선택을 통해 분류에 적합한 특징 집합을 선택하여 분류 성능의 향상에 초점을 두고 분류 실험을 수행하였다.

(표 1) 문서 분류 실험 데이터 구성

(Table 1) Construction of Document Classification Experiment data

범주	학습데이터	실험데이터	
		1	2
정치	1,000	100	100
경제	1,000	100	100
총계	2,000	200	200

4.2 실험 결과

4.2.1 특징 집합의 종류에 따른 분류 결과

본 연구에서 제안한 혼합 특징 집합의 효과를 검증하기 위해, 단일 모집단의 두 특징 집합과 본 연구에서 제안한 다중 모집단의 특징 집합을 각각 이용한 분류 실험을 수행하였다. 단일 모집단의 특징 집합으로는 인터넷 기사에서 추출한 단어들로 구성된 특징 집합과 LSA를 통한 변형 특징 집합을 사용하였으며, 다중 모집단의 특징 집합은 이 두 특징 집합을 혼합한 특징 집합을 사용하였다.

(표 2)는 학습 문서의 크기에 따른 세 가지 특징 집합을 이용한 특징 선택 후 분류 실험 결과인 분류 정확도 (accuracy)를 보인다. 선행 연구[14]가 원본 특징 집합의 크기보다 더 적은 크기의 특징 집합으로 유사한 분류 성능을 보였다면, 본 연구는 두 특징 집합을 혼합한 특징 집합을 이용하여 적은 크기의 특징 집합으로 기존 두 특징 집합을 이용한 분류 성능보다 향상된 분류 성능을 보인다. (표 2)에서 보이는 각 크기의 학습 문서에 따른 분류 성능을 보면, 제안한 특징 집합을 사용했을 경우 원본 특징 집합과 Semantic-based GAFS에 비해 적게는 1% 내외, 크게는 5% 이상 더 높은 분류 성능을 얻을 수 있었다.

(표 2) 세 가지 방법에 대한 문서 분류 정확도
(Table 2) Document Classification Accuracies of three methods

학습 문서 수	원본 특징 집합		Semantic-based GAFS		제안하는 방법	
	실험1	실험 2	실험 1	실험2	실험 1	실험2
500	74.9	70.5	75.9	77.3	89.4	85.7
1000	85.6	81.4	84	78.8	90.9	90.7
1500	87.8	83.1	88.8	87.3	92.2	89.3
2000	86.9	82.5	86	87	92	89
2500	89.2	84.3	87.2	81.1	92.7	89.2
3000	89.5	84.7	87.3	88.8	92.3	88.5
3500	90.5	87.5	91.6	90.8	92.2	89.5
4000	90.5	88	90.5	89.8	90.9	89.2

특히, (표 2)를 그래프로 표현한 (그림 5)와 (그림 6)에서 보이는 바와 같이 혼합 특징 집합을 사용 할 경우, 학습 문서의 크기가 작더라도 다른 단일 특징 집합보다 높은 분류 성능을 보임을 확인 할 수 있다. 또한, 학습 문서의 크기에 따른 분류 정확도의 차이가 5% 내로 적다. 이는 학습 문서의 양 대비 분류 효율이 높으며 크기에 상관 없이 일정한 분류 정확도를 가지는 것을 보인다.

(표 3)은 학습 문서의 크기에 따른 세 가지 특징 선택 방법을 이용한 분류 정밀도(precision)와 재현율(recall)을 보이며, 이로부터 본 연구에서 제안한 방법이 가지는 또 다른 특징을 확인하였다. (표 3)에서 굵은 글씨로 표시된 부분은 분류한 두 범주 각각에 대해 가장 크고 작은 차이를 보이는 정밀도와 재현율을 나타낸다. 학습 문서의 크

(그림 5) 세 가지 방법에 대한 문서 분류 정확도 I
(Figure 5) Document Classification Accuracies of three methods(I)

(그림 6) 세 가지 방법에 대한 문서 분류 정확도 II
(Figure 6) Document Classification Accuracies of three methods(II)

(표 3) 세 가지 방법에 대한 분류 정밀도 및 재현율 (Pol : 정치, Eco : 경제)
(Table 3) Classification Precisions and Recalls of three methods (Pol : Politics, Eco : Economy)

학습 문서 수	원본 특징 집합				Semantic-based GAFS				제안하는 방법			
	Precision		Recall		Precision		Recall		Precision		Recall	
	정치	경제	정치	경제	정치	경제	정치	경제	정치	경제	정치	경제
500	63.3	94.4	97.4	43.6	81.3	96.0	96.8	77.8	80.1	99.0	99.2	75.4
1000	74.4	94.1	95.8	67.0	87.9	90.1	90.4	87.6	94.8	94.0	94.0	94.8
1500	76.9	93.0	94.6	71.6	87.6	91.2	91.6	87.0	95.1	93.0	92.8	95.2
2000	75.5	94.8	96.2	68.8	85.4	89.4	90.0	84.6	93.4	93.4	93.4	93.4
2500	77.3	96.2	97.2	71.4	86.7	93.7	94.2	85.6	92.8	94.9	95.0	92.6
3000	78.1	95.3	96.4	73.0	86.6	93.8	94.4	85.4	94.9	93.3	93.2	95.0
3500	82.1	95.2	96.0	79.0	87.1	93.9	94.4	86.0	95.9	92.7	92.4	96.0
4000	82.9	95.0	95.8	80.2	88.2	94.2	94.6	87.4	94.5	92.6	92.4	94.6

기가 500인 경우 적은 학습 양으로 인해 모든 방법에서 20% 이상의 큰 차이를 보여 이를 제외한 나머지 학습 문서 크기에 대한 최대와 최소 차이 값을 선택하였다. 특징 선택 없이 원본 단어 집합만을 이용하였을 경우 정밀도는 최소 약 12%에서 최대 약 20% 사이의 차이를 보이며, 재현율은 최소 약 16%에서 최대 약 29%의 차이를 보인다. 또한, 선행 연구 방법인 Semantic-based FSGA은 정밀도에서 최소 약 2%에서 최대 약 7%의 차이를 보이며, 재현율은 최소 약 3%에서 최대 9%의 차이를 보인다. 이와 같은 결과는 분류 결과가 어느 한 쪽의 범주로 편향된 것을 의미한다. 하지만 본 연구에서 제안한 방법은 정밀도에서 최소 0%에서 최대 3%, 그리고 재현율에서 최소 0%에서 최대 4%로 앞의 두 방법에 비해 낮은 차이를 보인다. 즉, 분류 결과가 어느 한 쪽의 범주로 편향되지 않고 균등하게 분류가 이루어지는 것을 확인하였다.

(표 4) 선택 된 특징 집합의 크기
(Table 4) Sizes of Selected Feature Sets

특징 선택 방법	특징 집합 크기
원본 특징	6800
Semantic-based GAFS	134
제안하는 방법	142

또한, (표 4)에서 보이는 바와 같이 선행 연구[14]의 특징 선택 기법을 사용한 경우와 제안한 방법을 사용한 경우의 각 특징 집합의 크기는 유사하다. 혼합 특징 집합을 사용한 분류 시 유사한 크기의 특징 집합으로 더 나은 분류 효과를 얻었다. 이는 두 특징 집합의 정보를 모두 고려하여 반영한 혼합 특징 집합을 분류에 사용하였기 때문이다. 이 때 경계값으로 사용한 F 값은 원본 특징 집합의 경우 0.1, 변형 특징 집합은 0.001을 사용하였다. 각 F값은 학습데이터의 양(500~4000)과 반복적 학습을 통해 구해진, 특징 집합 크기 대비 가장 효율적인 분류 성능을 얻을 수 있는 특징 집합을 선택하는 경계값이다.

4.2.2. 특징 선택 경계값에 따른 분류 결과

특징 선택 방법은 분류에 적합한 특징 집합을 구성하기 위해 개개별 특징(Filter) 혹은 후보 특징 집합(Wrapper)에 대한 적합도 정도를 평가하였다. 이때 평가의 기준이 되는 경계값(혹은 기준값)을 정하는 것 역시 분류 성능에

매우 중요하다. 본 연구에서 제안한 방법은 필터 방법과 래퍼 방법 각각에 대한 경계값이 필요하다.

래퍼 방법의 경우 사전 연구에서 분류 정확도와 단순 반복 횟수를 이용한 유전알고리즘의 반복 제어하였을 시, 분류 정확도를 이용하였을 때 단순 반복 횟수를 제어한 경우보다 더 좋은 분류 정확도를 보였다. 하지만 성능 차이가 미비하기 때문에 반복 횟수를 이용한 특징 선택이 더 효율적임을 보였다.

(표 5) 학습 문서의 크기에 따른 각 특징 집합의 F 값
(Table 5) F-measures according to sizes of training data

학습 문서 수	원본 특징 F 값	LSA 특징 F 값
500	0.00858	0.00422
1000	0.00702	0.00395
1500	0.00655	0.00343
2000	0.00649	0.00367
2500	0.00629	0.00365
3000	0.00620	0.00361
3500	0.00603	0.00357
4000	0.00596	0.00361
평균	0.00664	0.00372

본 연구에서는 필터 방법의 경계값인 F 값에 따른 분류 성능에 초점을 맞추었다. 원본 특징 집합에 대한 F 값과 LSA를 통한 변형 특징 집합 각각에 대한 결과 F 값은 다음 (표 5)에서 보인다. 필터 방법의 경계값에 따른 분류 실험 결과는 (그림 7)에서 보인다. (그림 7)은 4000개의 학습 문서에 대한 특징 선택을 수행하여 선택된 특징 집합의 수와 그에 대응하는 분류 정확도를 보인다. 일정 경계값 이하부터는 분류 정확도가 일정하거나 떨어지는 것을 확인 할 수 있다. 이는 특정 경계값 이상의 특징들은 분류에 불필요한 특징으로, 선택 될 시 분류 성능에 악영향을 미치기 때문이다. 또한, 너무 높은 경계값 이상의 특징들 역시 분류 정확도가 낮아지는 것을 볼 수 있다. 이는 분류에 필요한 특징이 너무 높은 경계값으로 인해 제외되기 때문이다. 특징 집합의 크기 대비 성능으로 보았을 때, (그림 7)에 보인 바와 같이 특징 집합의 크기가 130~140개에 대한 F 값이 가장 분류에 효율적임을 알 수 있었다.

(그림 7) F값에 따른 선택 된 특징 집합의 크기와 분류 정확도
(Figure 7) Sizes and Classification Accuracies of Selected Feature Sets according to F-measures

5. 결 론

본 연구는 기존의 특징 선택 방법의 단일 모집단의 특징들에 대한 특징 선택이 아닌, 단어 집합과 단어 집합으로부터 LSA를 이용해 구한 의미 집합을 혼합한 특징 집합에 대해 특징 선택을 수행함으로써, 두 모집단의 특징들이 가지는 서로 다른 정보를 모두 고려하였다. 또한, 이로부터 최적의 특징 집합을 선택하기 위해 F값과 유전알고리즘을 이용한 하이브리드 특징 선택 방법을 제안하였다.

인터넷 기사의 분류 실험을 통해 두 가지 측면에서 제안한 방법의 효과를 검증하였다. 첫 번째로 학습 양에 상관없이 일정하고 높은 분류 정확도를 보였다. 제안한 방법은 학습 문서의 크기에 따라 최대 약 5%의 정확도 차이를 가지는 것을 보임으로써 학습 양에 상관없이 안정적인 분류 정확도를 보임을 확인하였다. 두 번째 효과로 어느 한 쪽의 범주로 분류 결과가 편향되지 않음을 분류 실험을 통해 확인하였다. 두 범주에 대한 정밀도와 재현율이 최대 3% 내의 낮은 차이를 보였다. 즉, 혼합 특징 집합을 이용할 경우 두 특징 집합의 상호 보완적인 관계를 통해 균형적인 분류 정확도를 얻을 수 있었다.

본 연구에서는 혼합 특징 집합을 두 가지 종류의 서로 다른 모집단을 가지는 특징 집합으로 구성하였지만, 또 다른 모집단의 특징 집합을 추가함으로써 더 많은 정보를 특징 선택 시 반영할 수 있다. 특징 집합을 혼합시키고 이에 대한 특징 선택을 수행함으로써 분류의 대상에 따라, 그리고 분류하고자 하는 범주의 종류에 따라 다양한 모집단의 특징 집합을 혼합하여 분류에 필요한 많은 정

보를 담은 특징 집합을 선택 할 수 있을 것이라 기대한다.

참 고 문 헌(Reference)

- [1] Guyon, I. Elisseeff, A., "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, Vol.3, pp.1157-1182, 2003.
- [2] Dasgupta, A. Drineas, P. Harb, B. Josifovski, V. Mahoney, M. W., "Feature Selection methods for Text Categorization", Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.230-239, 2007.
- [3] Landauer, T. K. Dumais, S. T., "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, Vol.104, No.2, pp.211-240, 1997.
- [4] Deerwester, S. C. Dumais, S. T. Landaner, T. K. Furnas, G. W. Harshman, R. A., "Indexing by latent semantic analysis", Journal of the American Society for Information Science, Vol.41, No.6, pp.391-407, 1990.
- [5] Chakraborti, S. Lothian, R. Wiratunga, N. Watt, S., "Sprinkling: Supervised Latent Semantic Indexing", Advances in Information Retrieval, pp.510-514, 2006.
- [6] Liu, H. Yu, L., "Toward Integrating Feature selection algorithm for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.4, pp.491-502, 2005.
- [7] Chen, C. M. Lee, H. M. Chang, V. J., "Two novel selection approaches for Web page Classification", Expert Systems with Application, Vol.36, No.1, pp.260-272, 2009.
- [8] Selamat, A. Omatu, S., "Web page Feature Selection and Classification using Neural Networks", Information Sciences, Vol.158, pp.69-88, 2004.
- [9] Yang, Y. Pedersen, J. O., "A Comparative Study on Feature Selection in Text Categorization", Proceedings of the 14th International Conference on Machine Learning(ICML '97), pp.412-420, 1997.
- [10] Peng, H. Long, F. Ding, C., "Feature selection Based on Mutual Information Criteria of Max-Dependency,

- Max-Relevance, and Min-Redundancy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No. 8, Aug. pp. 1226-1238, 2005.
- [11] John, G. Kohavi, R. Pflieger, K., “Irrelevant Feature and the Subset Selection Problem”, In Proceedings of 11th International Conference on Machine Learning, pp.121-129, 1994.
- [12] Luukka, P., “Feature selection using fuzzy entropy measures with similarity classifier”, Expert Systems with Applications, Vol.38, No.4, pp.4600-4607, 2011.
- [13] Gheyas, I. A. Smith, L. S., “Feature subset selection in large dimensionality domains”, Pattern Recognition, Vol.43, No.1, pp.5-13, 2010.
- [14] Kim, J. In, J. Chae, S., “Sementic-based Genetic Algorithm for Feature Selection”, Journal of Korean Society for Internet Information, Vol.13, No.4, pp.1-10, 2012
- [15] Liu, Y. N. Wang, G. Zhu, X. D., “Feature selection based on adaptive multi-population genetic algorithm”, Journal of Jilin University Engineering and Technology Edition, Vol.41, No.6, pp.1690-1693, 2011.
- [16] Sun, J. T. Chen, Z. Zeng, H. J. Lu, Y. C. Shi, C. Y. Ma, W. Y., “Supervised Latent Semantic Indexing for Document Categorization”, Fourth IEEE International Conference on Data Mining(ICDM '04), pp.535-538, 2004.

● 저 자 소 개 ●

인 주 호(Joo-ho In)

2006년 한국항공대학교 컴퓨터공학과(공학사)
 2008년 한국항공대학교 대학원 컴퓨터공학과(공학석사)
 2008년~현재 한국항공대학 대학원 컴퓨터공학과 박사과정
 관심분야 : 분산/병렬처리 시스템, 데이터마이닝
 E-mail : allpoyou@kau.ac.kr

김 정 호(Jung-ho Kim)

2008년 한국항공대학교 컴퓨터공학과(공학사)
 2010년 한국항공대학교 대학원 컴퓨터공학과(공학석사)
 2010년~현재 한국항공대학 대학원 컴퓨터공학과 박사과정
 관심분야 : 데이터마이닝, 패턴인식, 분산/병렬처리 시스템
 E-mail : natul2@kau.ac.kr

채 수 환(Soo-hoan Chae)

1973년 한국항공대학교 전자공학과(공학사)
 1985년 미국 Univ. of Alabama 대학원 전산학과(공학석사)
 1988년 미국 Univ. of Alabama 대학원 전기공학과(공학박사)
 1988년~현재 한국항공대학교 전자 및 정보통신공학부 교수
 관심분야 : 분산/병렬처리 시스템, 컴퓨터보안, 데이터마이닝
 E-mail : chae@kau.ac.kr