ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

의미 기반 유전 알고리즘을 사용한 특징 선택

Semantic-based Genetic Algorithm for Feature Selection

김 정 호* 인 주 호** 채 수 환***
Jung-ho Kim Joo-ho In Soo-hoan Chae

요 약

본 논문은 문서 분류의 전처리 단계인 특징 선택을 위해 의미를 고려한 최적의 특징 선택 방법을 제안한다. 특징 선택은 불필요한 특징을 제거하고 분류에 필요한 특징을 추출하는 작업으로 분류 작업에서 매우 중요한 역할을 한다. 특징 선택 기법으로 특징의 의미를 파악하여 특징을 선택하는 LSA(Latent Semantic Analysis) 기법을 사용하지만 기본 LSA는 분류 작업에 특성화 된 기법이 아니므로 지도적 학습을 통해 분류에 적합하도록 개선된 지도적 LSA를 사용한다. 지도적 LSA를 통해 선택된 특징들로부터 최적화 기법인 유전 알고리즘을 사용하여 더 최적의 특징들을 추출한다. 마지막으로, 추출한 특징들로 분류할 문서를 표현하고 SVM (Support Vector Machine)을 이용한 특정 분류기를 사용하여 분류를 수행하였다. 지도적 LSA를 통해 의미를 고려하고 유전 알고리즘을 통해 최적의 특징 집합을 찾음으로써 높은 분류 성능과 효율성을 보일 것이라 가정하였다. 인터넷 뉴스 기사를 대상으로 분류 실험을 수행한 결과 적은 수의 특징들로 높은 분류 성능을 확인할 수 있었다.

ABSTRACT

In this paper, an optimal feature selection method considering sematic of features, which is preprocess of document classification is proposed. The feature selection is very important part on classification, which is composed of removing redundant features and selecting essential features. LSA (Latent Semantic Analysis) for considering meaning of the features is adopted. However, a supervised LSA which is suitable method for classification problems is used because the basic LSA is not specialized for feature selection. We also apply GA (Genetic Algorithm) to the features, which are obtained from supervised LSA to select better feature subset. Finally, we project documents onto new selected feature subset and classify them using specific classifier, SVM (Support Vector Machine). It is expected to get high performance and efficiency of classification by selecting optimal feature subset using the proposed hybrid method of supervised LSA and GA. Its efficiency is proved through experiments using internet news classification with low features.

☞ keyword : 분류(Classification), 특징 선택(Feature Selection), 잠재적 의미 분석(Latent Semantic Analysis), 유전 알고리즘(Genetic Algorithm), 서포트 벡터 머신(Support Vector Machine)

1. 서 론

인터넷의 보급화로 인터넷 상에 대량의 문서가 범람함에 따라 자동 문서 분류의 필요성이 대두되었다. 과거 자동 문서 분류가 수작업으로 이루어 지던 분류 작업의 자동화였다면 현재는 얼마나 더 정확하고 빠르게 분류하는 것에 초점을 맞추고 있다[1]. 이를 위해 효율적인 분류 방법을 가진 분류기를 사용하거나 분류 대상을 표현하는 특징 집합을 최적화해야 한다. 본 연구에서는 분류에 최적화된 특징 집합을 찾는 특징 선택(feature selection)에 초점을 맞춘다.

특징 선택은 문서 분류뿐만 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 최적화 된 특징 집합은 분류 성능을 높이며 분류 속도 역시 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[2,3].

하지만 기존 특징 선택 연구들은 문서로부터 추출한 특징들을 정제하지 않고 사용한다. 문서를 표현하고자 하는 특징은 보통 문서 내에 포함된 단어이다. 단어를 특징으로 문서들을 표현할 경우 단어-문서 행렬은 문서의

* 정 회 원 : 한국항공대학교 대학원 컴퓨터공학과 박사과정
natul2@kau.ac.kr
** 정 회 원 : 한국항공대학교 대학원 컴퓨터공학과 박사과정
allpoyou@kau.ac.kr
*** 정 회 원 : 한국항공대학 전자 및 정보통신공학부 교수
chae@kau.ac.kr
[2012/01/18 투고 - 2012/01/30 심사(2012/05/02 2차) - 2012/06/05 심사완료]

http://dx.doi.org/10.7472/jksii.2012.13.4.1

크기가 클수록, 출현하는 단어의 종류가 많을 수록 매우 희소한 행렬(sparse matrix)이 된다. 단어-문서 행렬이 희소행렬일 경우 이로부터 제대로 된 정보를 얻을 수 없기 때문에 앞선 특징 선택 기법들을 사용하여 얻은 값을 신뢰할 수 없다.

이러한 희소 행렬 문제를 해결하면서 동시에 단어의 의미를 고려할 수 있는 방법으로 LSA가 있다. LSA는 단어-문서 행렬로부터 단어와 단어간의 공기 정보를 바탕으로 새로운 특징 집합을 생성한다. 생성된 특징 집합은 다음 세 가지 특징을 가진다. 1)단어들 간의 관계로부터 추출한 의미를 포함하며, 2)기존 특징 집합보다 축소된 크기를 가지며, 3)새로운 특징 집합으로 구성된 문서 행렬은 매우 밀집한 행렬(dense matrix)이다. LSA의 단어 간 공기 정보를 통한 의미 반영은 단어의 중의적인 성격을 파악하기 위한 동의어 연구[4]나 군집화(clustering) 연구[5]에서 입증되었다. 하지만 기본적인 LSA는 사전에 정의된 범주로의 분류에는 적합하지 못하다. 의미를 반영하는 특징 집합을 생성 시 분류할 범주에 대한 정보는 고려하지 않고 오직 단어들 간의 공기 정보만 고려하기 때문이다. 이러한 문제를 해결하기 위해 지도적 학습을 통한 지도적 LSA가 제안되었다[6,7]. 이는 지도적으로 분류할 범주에 대한 정보를 학습시키고 이를 반영한 새로운 특징 집합을 생성하는 것이다.

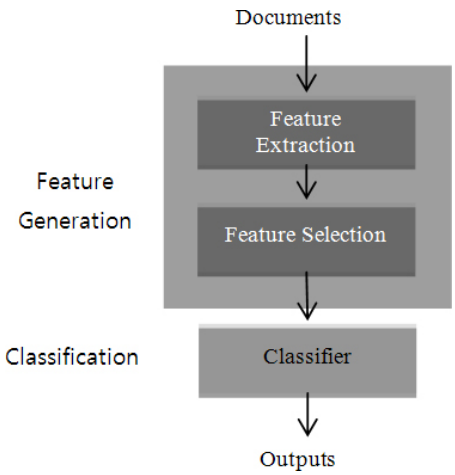
본 논문은 분류에 적합하며 특징의 의미를 고려한 특징 집합의 선택을 위해 지도적 LSA 기법 인 Sprinkling 기법과 이를 통해 구한 새로운 특징 집합으로부터 최적의 특징 집합을 구하는 유전알고리즘(genetic algorithm, GA)을 혼합한 의미기반 유전알고리즘 특징 선택 기법(Semantic-based GAFS)을 제안한다. 유전알고리즘은 전역 최적해(global optimal solution)을 탐색하는데 좋은 성능을 보이며 규모가 큰 최적화 문제에 적합한 방법으로[8], 규모가 큰 특징 집합으로부터 최적의 특징 집합(feature subset)을 탐색하는 문서 분류의 특징 선택 문제에 적합하다. 제안하는 방법의 성능을 검증하기 위해 인터넷 뉴스 문서 분류에 적용하여 실험하였다.

2. 특징 선택(feature selection)

문서 분류는 (그림 1)와 같이 크게 다음 세 가지 단계로 구성된다.

첫 번째 단계는 실제 문서를 단어-문서 행렬로 변환하는 것으로 특징 추출(feature extraction)이라 한다. 이는 문서를 표현하기 위한 특징을 생성하는 단계이기도 하다. 두 번째 단계는 추출 한 특징 중 불필요한 특징(redundant feature)를 제외한 부분 특징들만 선택하는 특징 선택(feature selection) 이다. 마지막 단계는 특징 선택 단계의 결과인 부분 특징 집합으로 문서를 분류하는 분류단계이다. 본 논문은 두 번째 단계인 특징 선택에 초점을 맞추었다.

특징 선택은 추출된 특징들 중 분류에 불필요한 특징들을 제거하고 분류에 필요한 최소한의 특징들을 선택한다. 최소한의 특징들을 선택함으로써, 분류 시간을 줄이고 정확도를 높일 수 있다. 데이터의 양이 기하급수적으로 증가하는 요즈음의 추세에선 효과적인 특징 선택은 매우 중요하다.



(그림 1) The System of Document Classification

최소한의 특징 선택을 위해, 특징 선택은 두 단계로 구성된다. 각각 전체 특징 집합으로부터 후보 부분 특징 집합을 탐색하는 단계, 후보 특징 집합을 평가하는 단계이다. 후보 특징 집합 탐색을 위한 기법들은 크게 완전탐색(complete search), 순차탐색(sequential search), 그리고 임의탐색(random search)가 있다[9].

탐색 결과인 후보 특징 집합은 분류에 적합한지 여부를 확인하기 위해 평가를 한다. 평가 기준은 특징 집합이 가지는 범주 간 분별력이다. 서로 다른 범주에 대해 큰 분별력을 가지는 특징 집합이 분류에 적합하다. 이러한 분별력 평가를 위한 접근 방법은 다음과 같이 크게 두 가지, Filter 방법, Wrappers 방법이 있다.

Filter 방법은 문서가 가지고 있는 정보를 바탕으로 후

보 특징 집합을 평가한다. 특징 집합으로 표현된 문서 벡터들 간의 거리나 IG(Information Gain)과 같은 측정치를 계산하여 이를 기준으로 특징의 선택 여부를 결정한다. 일반적인 측정치는 DF(Document Frequency), MI(Mutual Information), Chi, TS(Term Strength), Fuzzy rank 등이 있다[10-13]. Odds Ratio 같은 측정치도 존재하며 이를 개량한 EOR(Extended Odds Ratio), WOR(Weighted Odds Ratio)도 제안되었다[14]. Filter 방법은 독립된 단일 특징들의 가중치만을 가지고 특징을 선택하기 때문에 특징들 간의 관계를 고려하지 못하는 한계가 있다.

Wrapper 방법은 문서 분류에 사용할 분류기의 특성을 고려하기 위해 학습 데이터를 분류기로 분류하여 얻은 분류 성능으로 특징 집합을 평가한다. [15]는 Wrapper 방식의 특징 선택을 위해 linear SVM을 적용하였고, [16]는 Bayesian network를 적용하였다. Wrapper 방법은 모든 후보 특징 집합들에 대한 분류를 수행하기 때문에 Filter방법에 비해 계산 복잡도가 크다. 하지만 분류기의 특성이 고려되기 때문에 Filter 방법에 비해 분류 정확도가 높다. Wrapper 방법의 강점은 기존 Filter 방법과의 비교 연구를 통해 입증되었다[17]. 하지만 Wrapper 방법은 특징 학습 시 마다 매번 분류기의 학습이 병행되기 때문에 매우 비싼 작업이다. 이러한 이유로 최근에는 Filter 방식과 Wrapper 방식을 혼합한 Embedded 방식에 대한 연구가 진행되었다.[18,19] 기존 Filter 방식을 통해 후보 특징 집합들의 수를 줄이고, 줄어든 후보 특징 집합들로부터 Wrapper 방식을 이용하여 특징을 선택하는 방법으로 기존 Wrapper 방식의 느린 속도를 보안하면서 분류 성능을 보장한다. 하지만 여전히 특징의 중의적인 성격(동의어, 다의어 등)을 고려하지 않고 특징을 선택함으로 인해 생기는 성능 손실 문제는 여전히 존재한다. 본 연구는 단일 특징만을 독립적으로 고려하여 선택하지 않고 단어 간의 공기 정보를 통해 단어의 의미 관계를 반영한 특징 선택과 분류기의 특성을 고려하여 높은 분류 정확도를 얻기 위한 Wrapper 방법을 적용하였다.

3. LSA

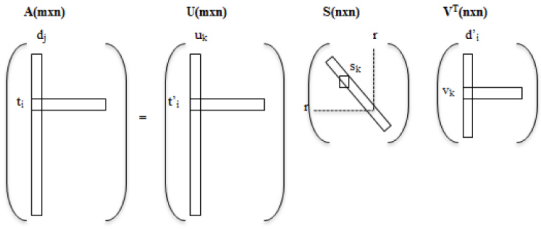
3.1 LSA 개념

LSA는 단어-문서 행렬로부터 단어-단어, 단어-문서 사이의 의미 관계를 구하는 기법이다. 의미 관계의 도출은 단어들 간의 공기(co-occurrence) 정보를 바탕으로 이루어지며, 의미 관계는 단어-문서 행렬로부터 구한 새로운 의미 공간(Semantic Space)에 표현된다. 의미공간은 행렬의 선형변환에 의거한 행렬 분해 기법인 SVD(Singular Value Decomposition)을 통해 얻어진다. SVD는 의미공간을 생성하는 것뿐만 아니라 불필요한 차원 제거를 통해 의미 공간의 차원을 축소시킨다. 이러한 차원 축소를 잡음 제거라고도 한다. 잡음을 제거하여 단어나 문서의 실제 의미를 더욱 정확하게 반영한다는 것에 그 의미가 있다.

SVD는 식 (1)과 같이 단어-문서 행렬(A_{mxn})을 세 개의 행렬로 분해하며, 각각 단어-단어의 상관관계를 나타내는 행렬(U_{mxn}), 문서-문서 상관관계를 나타내는 행렬(V_{nxn}), 그리고 두 행렬의 특이치(singular value)를 가지는 대각 행렬(S_{nxn})이다.

$$A = US V^T \quad (1)$$

행렬 U와 행렬 V의 행 벡터와 열 벡터는 정규 직교 기저 벡터로 구성 되며, 각각 단어-문서 행렬 A의 열공간과 행공간에 대한 기저 벡터이다.(그림 2)



(그림 2) Matrix Decompostion by SVD

즉, 행렬 U는 단어 간의 상관관계를 내포하는 새로운 의미 공간이고, 행렬 V는 문서 간의 상관관계를 내포하는 새로운 의미 공간이다. 그리고 행렬 S의 특이치는 행렬 U와 V의 기저 벡터가 의미 공간에서 가지는 강도를 의미하며, SVD의 차원 축소는 이 특이치를 기준으로 이루어진다. 특이치가 낮은 기저 벡터는 불필요한 차원으로 간주하여 제거하고 상위 k개의 특이치에 대응하는 열 벡터만 남긴 행렬 U_k와 V_k를 생성한다.

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

$$C = A U_k \quad (2)$$

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 정해둔 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.[6,7]

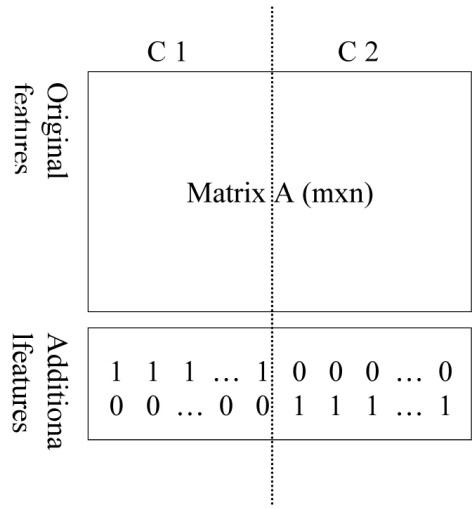
그 중 [6]은 기존 단어-문서 행렬에 범주 정보를 추가로 삽입하는 방법으로 SVD 행렬 분해 시 범주 정보 역시 포함 되도록 하는 Sprinkling 방법을 제안하였다. 그리고 [7]은 단어-단어 행렬인 U에서 서로 다른 범주간의 거리 차이를 기준으로 차원을 선택하는 방법을 제안하였다. 본 논문 역시 분류를 위해 지도적 학습이 추가된 LSA를 사용하며, 위 두 가지 방법 중 Sprinkling 을 사용하였다.

3.2 Sprinkling

Chakraborti가 제안한 sprinkling은 지도적 학습이 추가된 LSA 기법이다. Sprinkling은 단어-문서 행렬에 사전에 정의한 범주 정보를 삽입하고, 이로부터 의미공간을 생성하여 범주 정보가 의미공간에 포함되도록 한다. 이 의미공간은 동일 범주에 속하는 문서간의 동질성을 높이고 서로 다른 범주의 문서끼리의 이질성을 높인다.

(그림 3)은 sprinkling의 지도적 학습 방법을 보인다. 분류 하고자 하는 두 범주 C1과 C2가 있을 때, 단어-문서 행렬에 두 범주에 대한 행을 추가한다.

범주 C1에 속하는 문서의 경우 추가 된 첫 번째 행의 값을 1로, 범주 C2에 속하는 문서의 경우 두 번째 행의 값을 1로 설정한다. 행렬 분해 시, 추가된 범주 값으로 동일 범주의 단어들이 의미적으로 연결되기 때문에 의미공간에 범주 정보를 포함시킬 수 있다. 결과적으로 의미공간에 문서를 투영할 시 동일 범주끼리 더욱 가깝고 다른 범주끼리 멀어진다. 즉, 분류 범주 정보를 지도적으로 추가하여 서로 다른 범주간의 차별성을 증가시켜 분류 작업을 용이하게 한다.



(그림 3) Sprinkled Matrix

4. Genetic Algorithm

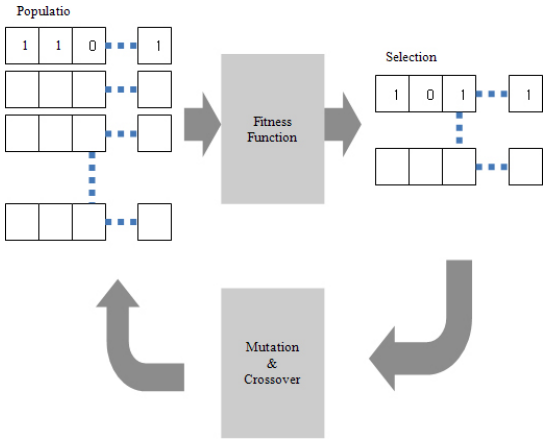
유전 알고리즘은 생물의 유전 진화의 메커니즘을 기반으로 한 공학 모델로 최적화 탐색 문제에 적합한 알고리즘이다. 유전 진화 메커니즘을 그대로 모델링 하였기 때문에 실제 유전자 변이(mutation), 상속(inheritance), 교배(crossover)를 적용하여 최적의 답을 찾는다.

유전 알고리즘은 풀고자 하는 문제의 가능한 여러 답들을 염색체(chromosome)로 나타낸다. 이때 염색체는 이진수열로 표현되는 경우가 보통이고 문제에 따라 더 복잡하게 구성될 수 있다. 구성한 염색체들은 각각 얼마나 문제의 답에 최적화 되었는지 평가하는 평가함수(fitness function)를 통해 평가 된다. 평가 결과를 바탕으로 높은 평가를 받은 유전자들 사이의 교배, 상속, 그리고 변이를 수행하여 더 좋은 염색체(최적의 답)를 찾는다. 이는 좋은 염색체 간의 교배, 상속, 그리고 변이를 할 시 더 좋은 염색체가 생성 될 것이라는 building block 가설을 따른다.

유전 알고리즘은 (그림 4)과 같은 단계를 반복적으로 수행한다.

염색체의 구성은 크게 세 가지 코딩 방식을 따른다. 이 중 이진 코딩을 가장 많이 사용하며, 상황과 목적에 따라 다른 코딩 방식을 사용한다. 이 세 가지 외에도 필요에 따라 적절한 코딩 방식을 만들어 사용하기도 한다.

1) 이진 코딩(binary encoding)
2) 실수 코딩(real encoding)
3) 심볼릭 코딩(symbolic encoding)



(그림 4) The Task Cycle of Genetic Algorithm

유전 알고리즘 시 주요 주의사항은 반복적인 해의 탐색 중 지역 최적점(local maximum)에 빠지지 않도록 하는 것이다. 이를 위해 효과적인 변이와 교배 메커니즘을 정의해야 한다. 또한, 찾고자 하는 최적의 답에 도달 할 수 있도록 평가 함수의 평가 기준을 명확하고 정확하게 정의해야 한다.

5. Semantic-based GAFS

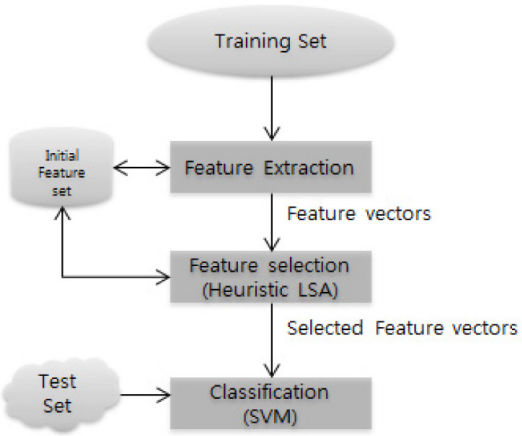
제안하는 특징 선택 방법인 Semantic-based GAFS는 문서 분류에 필수적인 최소의 특징들을 선택하여 최적의 특징 집합을 생성하는 것이다. 최적의 특징 집합이라 함은 문서를 정확하게 표현하고 최소한의 특징들로 구성됨을 의미한다. 분류의 성능은 문서를 표현하는 특징의 질에 많은 영향을 받기 때문에 효과적인 특징 선택 작업이 필수적이다.

본 논문에서 제안하는 특징 추출 방법은 크게 두 가지 단계로 구성된다.

1) 의미 분석 및 잡음 제거를 위한 지도적 LSA
2) 최적의 특징 집합을 찾기 위한 유전 알고리즘

첫 번째 단계로 지도적 학습 LSA를 사용함으로써 단어의 의미와 분류 범주 정보를 내포하는 새로운 특징 집합을 선택한다. 이를 통해 특징 선택 시 단어 간의 관계를 고려할 수 있다. 또한, 특이치를 기준으로 특징 집합의 크기를 축소하여 의미 없는 특징을 선택 대상에서 제외 한다. 특징 집합을 축소하였기 때문에 최적의 특징 집합을 탐색하는 탐색 시간이 줄어든다. 다음 단계로 LSA를 통해 선택 된 특징 집합으로부터 분류에 사용될 최적의 특징들을 선택하기 위해 유전 알고리즘을 사용한다. 유전 알고리즘은 최적화 문제(optimization problem)를 풀기 위한 방법으로 규모가 큰 특징 집합으로부터 최적의 특징을 찾는 문제에 적합하다. 또한 평가 기준 및 유전 연산의 적절한 조작에 따라 기존 특징 선택 기법보다 더욱 빠르게 최적의 해에 수렴하기 때문에 특징 집합 선택 문제에 적합하다.

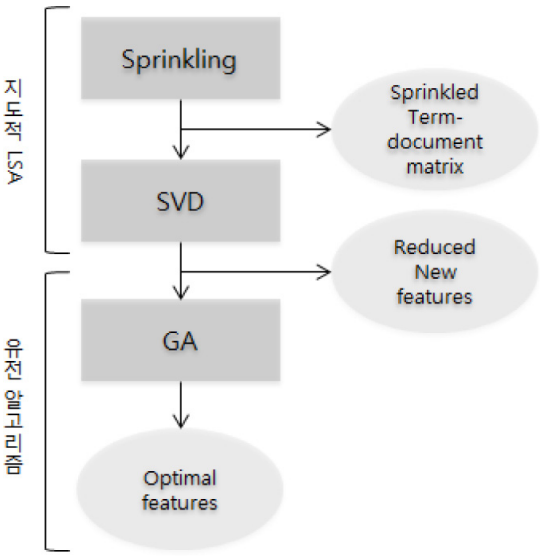
본 연구에서 제안하는 특징 선택 방법을 적용한 분류 시스템의 구성은 다음 (그림 5)와 같다. 크게 학습 단계인 특징 선택과 실제 분류를 수행하는 분류단계로 구성된다. 학습 문서로부터 최적의 특징 집합을 생성하고 이를 실제 실험 데이터에 적용하여 분류 작업을 수행한다.



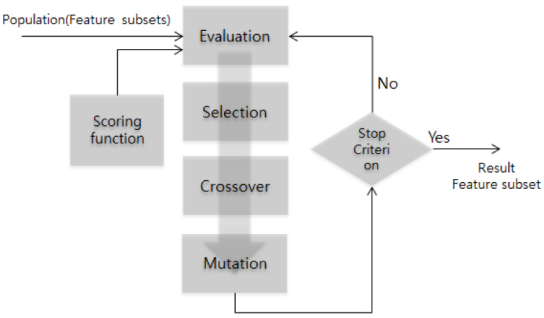
(그림 5) The System of Document Classification

(그림 6)은 학습 단계에서 수행되는 작업 과정을 자세히 보인다.

초기 단어-문서 행렬이 입력으로 들어오면 먼저 지도적 LSA 작업을 수행한다. sprinkling을 통해 단어-문서 행렬에 범주 정보를 담은 새로운 행을 추가한다. 범주 정보가 추가된 단어-문서 행렬은 SVD 단계를 거쳐 축소된 새로운 특징 집합을 생성한다. 이때 특징 집합은 SVD의 결과 중 행렬 U의 행 벡터들이며, 이 특징 집합은 단어 간의 공기 정보를 바탕으로 도출해 낸 단어의 의미 정보를 내포하고 있다. 지도적 LSA 단계의 결과인 축소된 특징 집합에 최적의 특징을 선택하기 위한 유전 알고리즘을 적용한다. (그림 7)은 유전 알고리즘을 적용한 특징 선택 과정을 보인다.



(그림 6) Semantic-based GA Feature Selection



(그림 7) The task cycle of GA

유전 알고리즘의 첫 단계는 초기 염색체 집합을 임의로 생성하는 것이다. 염색체의 각 유전자는 개별 특징을 의미하며 그 값이 1인 경우 해당 특징이 선택되었음을 의미한다. 초기 염색체의 유전자 즉, 특징은 임의로 선택되며 각 염색체 별로 100개의 특징을 선택한다.

초기 염색체 집합의 구성이 완료되면 평가 함수를 통해 각 염색체를 평가한다. 염색체의 평가는 분류 시스템의 분류기를 사용하여 이루어진다. 본 연구에서 사용하는 분류기는 SVM이며, SVM을 사용하여 각 염색체가 선택한 특징 집합으로 학습 문서를 표현하고 이를 분류한 결과를 평가 기준으로 사용한다.

평가함수를 거쳐 모든 초기 염색체의 평가가 완료된 후, 평가 점수를 바탕으로 상위 50%의 염색체를 제외한 나머지 하위 염색체는 제거한다. 이러한 작업을 유전연산자 중 선택(selection)이라 한다. 선택 된 상위 50%의 염색체는 두 가지 유전자 연산을 거쳐 새로운 염색체를 생성한다. 먼저 임의의 두 염색체 간의 1)교배를 수행하고, 2)그 결과로 변이 작업을 수행한다. 변이 연산은 지역 최적점을 피하기 위한 중요한 단계이므로 되도록 임의의 특징들을 선택하여 그 값을 반전(1->1 ‖ 0->1)시켰다.

유전연산을 통해 새롭게 구성 된 염색체 집합은 다시 평가 함수를 거쳐 평가되고, 그 결과로 선택, 교배, 변이 연산을 반복 수행한다. 유전 알고리즘의 이러한 반복 작업은 특정 조건을 만족 할 때까지 반복된다. 본 연구에서는 평가 점수가 90%를 초과하는 염색체가 존재 할 시 반복 작업을 종료한다.

유전 알고리즘의 결과로 구한 최종 염색체는 최적의 특징 집합이며 이를 사용하여 분류할 실험 문서를 표현하고 분류를 수행한다.

6. 실험 및 결과

학습 및 실험 데이터는 조선 일보의 인터넷 기사를 사용하였다. 분류 할 범주는 '정치'와 '경제' 이며, 이 두 범주의 문서들은 서로 유사하거나 중복되는 단어가 많이 발생하므로 두 범주의 경계가 모호하다. 본 연구의 실험은 이런 모호성에도 두 범주를 구별해주는 최적의 특징을 선택하여 높은 분류 성능을 얻는 것이 주 목적이다.

실험 데이터의 구성은 다음과 같이 구성하였다. 학습 문서로 총 4500개의 인터넷 기사를 사용하였으며, 각 범주 별로 2250개 씩 구성하였다. 이 중 특징을 추출하여 선택하기 위한 데이터로 4500개 중 4000개를 사용하였고, 남은 500개는 유전 알고리즘의 평가 대상으로 사용하였다. 실험 문서는 총 1000개의 인터넷 기사를 사용하였고, 각 범주 별로 500개 씩 구성하였다.

위와 같이 구성한 학습 및 실험 데이터를 가지고 본 연구는 제안한 특징 선택 방법의 효과를 검증하기 위해 다음 3 가지 관점으로 실험을 하였다.

1) 선택된 특징 집합의 크기 및 평가 점수
2) 학습 문서의 개수에 따른 성능 비교
3) 다른 특징 선택 기법들과의 성능 비교

6.1 선택된 특징 집합의 크기 및 평가 점수

특징 선택은 최적의 특징을 찾는 것이 목적이므로, 선

택 된 특징 집합의 크기가 얼마만큼 최소화 되었으며 그 점수가 얼마만큼 높은지를 확인하였다.

다음 (표 1)은 특징 집합의 크기와 그에 따른 평가 점수를 보인다. 본 논문에서 제안한 특징 선택 기법을 사용하여 선택 된 상위 10개의 특징 집합에 대한 크기와 학습 정확도, 그리고 평가 점수 및 분류 정확도를 보인다. 상위 평가 점수를 받은 특징 집합은 하위 특징 집합에 비해 특징 집합 크기 대비 분류 정확도가 높은 것을 확인 할 수 있었다. 또한, 실험 결과에서 주목할 점은 불필요한 특징의 존재 여부에 따라 성능 차이가 난다는 것이다. <표 1>에서 1 순위의 특징 집합의 크기가 133인데에 비해 2 순위의 특징 집합의 크기는 146개로 더 크다. 하지만 1 순위의 특징 집합이 크기 대비 분류 정확도가 높다. 이는 2순위 특징 집합에 불필요한 특징들이 존재하며, 이러한 점이 분류 성능에 악영향을 미친다는 것을 알 수 있다. 반대로 분류에 필요한 특징이 선택되지 못해 분류 성능이 낮은 경우도 있다. 7 순위의 특징 집합은 특징 집합 중 가장 작은 크기를 가지나 평가 점수 및 분류 정확도가 상위 크기가 더 큰 특징 집합에 비해 낮다. 이는 분류에 필요한, 분별력이 큰 특징이 선택되지 못한 경우이다.

(표 1) 특징 집합 크기에 따른 평가 점수 및 분류 정확도

순위	특징집합 크기	학습 정확도(%)	평가 점수(F)	분류 정확도(%)
1	**133**	**92.7%**	**0.9176**	**91.6%**
2	146	92.9%	0.9168	90.8%
3	133	90.7%	0.9016	90.5%
4	129	90.6%	0.9015	91.1%
5	132	90.0%	0.8962	88.2%
6	129	89.8%	0.8951	85.6%
7	120	89.5%	0.8944	81.2%
8	129	89.7%	0.8943	84.6%
9	128	89.5%	0.8929	86.2%
10	126	89.4%	0.8925	85.8%

본 실험에서 반복 작업의 종료 기준은 평가 점수 90% 이상이며, 위 결과는 평가 점수에 이르기 까지 10회 내외의 반복 작업이 수행된 결과이다. 평가 점수의 기준을 더욱 높인다면 더 높은 분류 성능을 기대할 수 있다.

6.2 학습 문서의 개수에 따른 성능 비교

특징 선택을 위해 학습 문서의 구성을 500개 단위(정치 250개, 경제 250개)로 구성하였다. 각각 1000개, 1500개, 2000개, 2500개, 3000개, 3500개, 4000개로 학습 문서를 구성하여 학습 단계를 수행하였다. 실험 데이터는 500개(정치 250개, 경제 250개)의 문서로 구성하였으며 총 5개의 실험 데이터를 사용하여 실험하였다. 특징 집합을 선택할 학습 문서의 크기가 커질수록 더욱 가치 있는 특징들을 선택 할 수 있기 때문에 학습 문서의 크기에 비례하여 분류 정확도가 향상 되었다. <표 2>에서 보이는 바와 같이, 학습 문서의 수에 비례하여 분류 정확도가 향상되는 것을 확인 할 수 있다.

하지만 학습 문서의 크기에 반비례로 분류 정확도가 떨어지는 경우도 있으나, 이는 10% 미만의 매우 미비한 차이이므로 전체적인 정확도 향상에는 크게 영향을 미치지 않는다.

6.3 다른 특징 선택 기법들과의 성능 비교

본 논문에서 제안한 특징 선택 기법의 성능을 검증하기 위해 기존 특징 선택 기법과의 분류 성능을 비교 분석하였다. 대표적인 특징 선택 기법인 SFFS(Sequential Forward Floating Search)과 분류 성능을 비교하였다. SFFS는 기존 다른 특징 선택 기법들 중 가장 성능이 좋은 특징 선택 기법이며, 이는 실험을 통해 증명된 바가 있다[20].

SFFS를 포함하여 총 3가지 특징 선택 기법과 비교하였으며 각각 다음과 같다.

1) No Feature Selection Methods
2) SFFS
3) LSA + SFFS

(표 3)은 위 3 가지 경우에 대한 분류 결과를 보인다. 실험은 4000개의 학습 문서로 학습을 수행하고 서로 다른 5가지의 1000개의 실험 데이터로 분류 실험을 수행한 결과이며, 각 비교 대상 별 평균 정확도를 보인다. 특징 선택 기법을 사용하지 않은 경우 평균 약 88%의 분류 정확도를 보인다. 하지만 SFFS를 사용하였을 경우 특징집합의 크기가 평균 130개이며 정확도가 약 86%로 특징 선택 기법을 사용하지 않은 경우와 크게 차이 나지 않는다. 6800개의 기존 특징 개수보다 훨씬 적은 130개의 특징들로만 유사한 분류 정확도를 얻을 수 있다. 본 논문에서 제안한 Semantic-based GAFS의 경우 평균 134개의 특징들을 선택하며 그 성능이 약 90%를 보였다. 다른 세 경우에 비교해 본다면 특징 집합의 크기 대비 분류 정확도

(표 2) 학습 문서 개수에 따른 분류 정확도

실험 데이터	학습 데이터						
	1000	1500	2000	2500	3000	3500	4000
1	0.759	0.84	0.86	0.872	0.873	0.905	0.916
2	0.773	0.788	0.87	0.811	0.888	0.898	0.908
3	0.756	0.774	0.788	0.863	0.838	0.861	0.905
4	0.745	0.778	0.822	0.806	0.879	0.812	0.916
5	0.88	0.87	0.802	0.858	0.843	0.862	0.882

가 높은 것을 확인 할 수 있다. 즉, 본 논문에서 의도한 지도적 LSA의 특징의 의미 분석과 유전 알고리즘의 최적의 특징 집합 선택을 통한 성능 향상을 확인 할 수 있다.

또한, SFFS와 Semantic-based GAFS의 가장 큰 차이점은 SFFS는 특징 집합의 크기를 사용자가 직접 정해주어야 하지만 Semantic-based GAFS는 특징 집합의 크기를 자동으로 정한다는 점이다. 최적의 특징 집합의 크기를 미리 예측하여 정하기는 매우 힘들며 그 크기가 분류하는 대상에 따라 매우 가변적이기 때문에 지도적으로 그 크기를 설정하는 것은 매우 비효율적이다. 자동으로, 경험적으로 최적의 특징 집합을 탐색할 경우 다양한 분야의 분류에 적용하기 좋다.

(표 3) 특징 선택 기법 비교 실험

특징선택 기법	특징집합 크기	분류 정확도(%)
No FS	6800	0.882
SFFS	130	0.865
Semantic-based GAFS	134	0.9054

시간적인 측면으로도 제안한 방법은 6.1의 실험 결과에서 보이는 바와 같이 적은 횟수의 반복 작업을 수행한다. [21]에서 제시한 바와 같이 특징 선택 문제에서 초기 특징 집합의 크기가 N일 때, 유전 알고리즘은 한 번의 탐색 시 $\Theta(N)$의 시간 복잡도(time complexity)를 가지며 SFFS는 $O(2^N)$의 시간 복잡도를 가지기 때문이다. 특히 유전 알고리즘은 특징 집합의 규모가 큰($N>50$) 경우 시간 측면의 성능이 높기 때문이다. 사전 작업인 지도적 LSA의 복잡도는 결과인 후보 특징 집합의 크기가 M일 경우 $O(NM^2)$으로 추출할 후보 집합의 크기에 영향을 많이 받지만 최종 특징 집합을 찾는 반복 작업 전에 단 한 번 수행되므로 전체 수행 시간에 큰 비중을 차지하지 않는다. 본 논문의 실험 결과를 통해서도 최적의 특징 집합을 찾기 위해 적은 반복작업을 수행한다는 것을 확인 할 수 있다.

7. 결 론

문서 분류를 위한 특징 선택 기법인 Semantic-based GAFS를 제안하였다. 지도적 LSA를 사용하여 특징들의 의미를 고려하고, 최적화 기법인 유전알고리즘을 적용하여 최적의 특징 집합을 찾았다. 지도적 LSA는 기존 매우 희소한 행렬인 단어-문서 행렬을 밀집 행렬로 변환함으로써 출현한 단어의 수가 낮아 분류가 제대로 되지 못하는 문서의 수를 줄여 분류 성능을 높였다. 그리고 지도적 학습을 통해 분류에 적합한, 분별력이 큰 특징들을 생성하였다. 마지막으로 유전알고리즘을 적용하여 지도적 LSA를 통해 생성된 특징 집합으로부터 최적의 특징 집합을 탐색하였다.

그 결과, 기존 특징 선택 기법보다 더욱 최적화 된 특징 집합을 찾았다. 또한, 기존 특징들의 의미가 반영된 특징 집합을 생성함으로써 기존 특징 선택 기법과 비교하여 특징 집합의 크기 대비 높은 분류 정확도를 얻었다.

본 연구가 빠르고 정확한 문서 자동분류에 기여할 것을 기대한다.

Acknowledgements.

본 연구는 2011년도 한국항공대학교 교내 연구비 지원을 받아 이루어졌음

참 고 문 헌

[1] X. Qi, B. D. Davison, “Web page classification: Features and Algorithms,” ACM Computing Surveys(CSUR), Vol. 41, No. 2, Feb. 2009, pp. 1-31.

[2] I. Guyon, A. Elisseeff, “An Introduction to Variable and Feature Selection,” Journal of Machine Learning Research, Vol. 3, Jan. 2003, pp. 1157-1182.

[3] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, M. W. Mahnoney, “Feature Selection methods for Text Categorization,” Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 230-239.

[4] Landauer, T. K., S. T. Dumais, “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge,” Psychological Review, Vol. 104, No. 2, Apr. 1997, pp. 211-240.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landaner, G. W. Furnas, R. A. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, Vol. 41, No. 6, 1990, pp. 391-407.

[6] S. Chakraborti, R. Lothian, N. Wiratunga, S. Watt, “Sprinkling: Supervised Latent Semantic Indexing,” Advances in Information Retrieval, 2006, pp. 510-514.

[7] J. T. Sun, Z. Chen, H. J. Zeng, Y. C. Lu, C. Y. Shi, W. Y. Ma, “Supervised Latent Semantic Indexing for Document Categorization,” Fourth IEEE International Conference on Data Mining(ICDM ’04), Nov. 2004, pp. 535-538.

[8] L. S. Oliveira, N. Benahmed, R. Sabourin, F. Bortolozzi, C. Y. Suen, “Feature Subset Selection Using Genetic Algorithms for Handwritten Digit Recognition,” Proceeding SIBGRAPI ’01 Proceedings of the 14th Brazilian Symposium on Computer Graphics and Image Processing, 2001, pp.362-370

[9] H. Liu, L. Yu, “Toward Integrating Feature selection algorithm for Classification and Clustering,” IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, 2005, pp. 491-502.

[10] C. M. Chen, H. M. Lee, Y. J. Chang, “Tow novel feature selection approaches for Web page Classification,” Expert Systems with Application, Vol. 36, No. 1, Jan. 2009, pp. 260-272.

[11] A. Selamat, S. Omatu, “Web page Feature Selection and Classification using Neural Networks,” Information Sciences, Vol. 158, Jan. 2004, pp. 69-88.

[12] Y. Yang, J. O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” Proceedings of the 14th International Conference on Machine Learning(ICML ’97), Jul. 1997, pp. 412-420.

[13] H. Peng, F. Long, C. Ding, “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. 2005, pp. 1226-1238.

[14] J. Cheng, H. Huang, S. Tian, “Feature Selection for Text Classification with Naïve Byes,” Expert Systems with Application, Vol. 36, No. 3, Apr. 2009, pp. 5432-5435.

[15] D. Mladenic, J. Brank, M. Grobelnik, N. Milic-Frayling, “Feature selection using Linear Classification weights: Interaction with Classification models,” Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 2004, pp. 234-241.

[16] I. Inza, P. Larranaga, R. Etxeberria, B. Sierra, “Feature Subset Selection by Bayesian network-based Optimization,” Artificial Intelligence, Vol. 123, No. 1-2, 2000, pp. 157-184.

[17] G. John, R. Kohavi, K. Pfleger, “Irrelevant Feature and the Subset Selection Problem,” In Proceedings of 11th International Conference on Machine Learning, 1994, pp. 121-129.

[18] P. Luukka, “Feature selection using fuzzy entropy measures with similarity classifier,” Expert Systems with Applications, Vol. 38, No. 4, Apr. 2011, pp. 4600-4607.

[19] I. A. Gheyas, L. S. Smith, “Feature subset selection in large dimensionality domains,” Pattern Recognition, Vol. 43, No. 1, Jan. 2010, pp. 5-13.
[20] J. Hua, W. D. Tembe, E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” Pattern Recognition, Vol. 42, No. 3, Mar. 2009, pp.409-424.
[21] M. Kudo, J. Sklansky, “Comparison of Algorithms that Select Features for Pattern Classifiers,” Pattern Recognition, Vol. 33, No. 1, 2000, pp. 25-41.

◐ 저 자 소 개 ◑

김 정 호 (Jung-ho Kim)
2008년 한국항공대학교 컴퓨터정보공학과 졸업(학사)
2010년 한국항공대학교 대학원 컴퓨터공학과 졸업(석사)
2010년~현재 한국항공대학교 대학원 컴퓨터공학과 박사과정
관심분야 : 데이터마이닝, 패턴인식, 분산/병렬처리 시스템
E-mail : natul2@kau.ac.kr

인 주 호 (Joo-ho In)
2006년 한국항공대학교 컴퓨터정보공학과 졸업(학사)
2008년 한국항공대학교 대학원 컴퓨터공학과 졸업(석사)
2008년~현재 한국항공대학교 대학원 컴퓨터공학과 박사과정
관심분야 : 분산/병렬처리 시스템, 데이터마이닝
E-mail : allpoyou@kau.ac.kr

채 수 환 (Soo-hoan Chae)
1973년 한국항공대학교 전자공학과 졸업(학사)
1985년 미국 Univ. of Alabama 대학원 전산학과 졸업(석사)
1988년 미국 Univ. of Alabama 대학원 전기공학과 졸업(박사)
1988년~현재 한국항공대학 전자 및 정보통신공학부 교수
관심분야 : 분산/병렬처리 시스템, 컴퓨터보안, 데이터마이닝
E-mail : chae@kau.ac.kr