특허 분류를 위한 효과적인 자질 선택

정하용° 황금하 신사임 최기선
한국과학기술원 전자전산학과 전산학전공
{hymanse°,hgh, miror}@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Effective Feature Selection for Patent Classification

Ha-Yong Jung° Jin-Xia Huang, Sa-Im Shin, Key-Sun Choi
Dept. of Computer Science, KAIST

요 약

자질 선택은 문서 분류와 같이 많은 자질을 사용하는 지도식 기계학습에 관한 연구에서 날로 중요성이 커지고 있다. 특히 특허문서 분류와 같은 작업은 기존의 문서 분류보다도 훨씬 많은 자질과 분류 범주를 가지기 때문에 전체 문서의 특징을 드러내는 적절한 부분집합을 선택해 학습하는 것이 절실하다. 전통적인 자질선택 방법은 필터라는 방법으로서 빠르지만 임계값을 정하기가 어렵다는 문제가 있다. 한편 최근에 많이 연구되는 래퍼는 일반적으로 필터보다 좋은 성능을 보이지만 자질의 개수가 많을수록 시간이 오래 걸린다는 단점이 있다. 본 연구에서는 필터와 래퍼를 상호 보완적으로 결합하여 최적의 필터를 자동적으로 찾는 래퍼를 제안한다. 실험 결과, 제안한 방법이 효과적으로 자질 집합을 선택하는 것을 확인할 수 있었다.

1. 서 론

지도식 기계학습 방법에는 가장 기본적인 베이즈 학습부터 시작하여 신경회로망, 최대 엔트로피 모델, 서포트 벡터 머신등 다양한 방법들이 존재한다. 이런 다양한 지도식 기계학습 방법들은 보통 수십여 개 정도의 자질들과 그에 해당하는 정답간의 통계적 정보를 다양한 방법으로 이용해, 새로운 자질들이 주어졌을 때 확률적으로 가장 근사한 해를 구한다. 지도식 기계학습에 사용되는 자질들은 학습에 있어 가장 중요한 부분이지만, 모든 자질들이 학습에 영향을 끼치는 것은 아니다. 예를 들어 언제나 값이 일정한 자질은 학습에 영향을 주지 못하고, 정답과 전혀 연관이 없는 자질은 학습의 입장에서 무작위적으로 비춰져 오히려 학습 성능을 떨어트릴 수도 있다. 따라서 학습에 큰 영향을 끼치는 자질을 선택해 그 자질들만을 이용해 학습한다면 학습 성능은 높이고 학습 시간은 줄일 수 있을 것이다.

자질 선택에 관한 연구는 크게 필터와 래퍼라는 두 가지 접근 방법이 있다. 필터는 전통적인 접근 방법으로서 자질들의 통계치를 이용해 일정 임계값 이상의 자질만을 선택하는 방법이고, 래퍼는 기계 학습 모델 자체를 블랙박스로 이용해 모델에서 최대값을 가지는 자질을 선택하는 실용적 접근 방법이다.

본 논문에서 제안하는 접근 방법은 필터와 래퍼를 결합한 방법이다. 래퍼는 일반적으로 필터에 비해서 좋은 성능을 보이지만 시간이 너무 오래 걸린다는 단점이 있다. 우리는 이를 보완하기 위해 필터를 적용한 자질 집합 중 최적의 집합을 찾는 데에 래퍼를 사용했다. 즉, 기존의 래퍼가 최적의 자질을 찾는 래퍼 였다면, 본 연구의 래퍼는 최적의 필터를 찾는 래퍼이다.

실험은 특허 문서를 분류하는 작업에서 최고의 정확률을 가지는 자질 집합을 제안한 방법으로 찾도록 하였다. 특허 분류는 자질이 대단히 많고 분류 범주 또한 무척 상세해서 효과적인 자질의 선택이 성능에 큰 영향을 끼칠 수 있다. 실험 결과 다양한 필터들 중 가장 좋은 성능을 보이는 필터를 자동적으로 선택할 수 있었다.

2. 관 련 연 구

2.1. 자질 선택

전술한 바와 같이 효과적인 기계학습을 위해 사용하는 자질 선택 방법에는 크게 두 가지가 있는데, 필터와 래퍼가 그것이다. 우선 필터는 기존의 자질선택 연구들이 사용해 오던 전통적인 방법으로서 자질들의 분포에서 드러나는 통계적 수치들을 이용하여 자질에 해당하는 통계치가 일정한 임계값 이상 되는 자질들만을 선택하는 방법이다.[1,4,5] 필터는 가능한 모든 자질들에서 부분 집합을 선택하는 일종의 전처리로서, 실제로 훈련과 예측에 사용하게 될 기계학습 방법과는 완전히 독립적이다. 필터는 일종의 연관성 점수를 각각의 자질에 부여하는 것이라고 생각할 수 있는데, i번째 자질과 정답과의 연관성 점수를 S(i)라 하면, S(i)가 일정 임계값 이상인 모든 자질을 선택하는 방법이라 할 수 있다. 이처럼 연관성 점수 S(i)는 개개의 자질에서 정의되기 때문에 각각의 S(i)는 완전히 독립적이다. 그리고 필터는 자질 집합의 통계치를 이용해 연관성 점수 S(i)를 구하기만 하면 되므로 빠르다는 장점이 있다. 필터를 설계할 때의 문제는 어떤 통계치를 사용하고, 임계값을 어느 정도로 설정하느냐인데 이 설정에 따라 필터의 성능이 결정되게 된다.

래퍼는 최근에 응용되기 시작한 방법으로서 훈련을 통해 생성된 기계학습 모델 자체를 일종의 블랙박스로 사용하여 자질들을 평가하는 방법이다. [1,6,7] 앞서 언급했던 필터가 실제로 훈련과 예측에 사용할 기계학습 방법과 완전히 독립적이었던 것과 반대로 래퍼는 특정 기계학습 방법에 완전히 의존적이다. 즉, 하나의 래퍼에서 선택한 자질 집합이 기계학습 방법이 달라지면 전혀 좋지 못한 자질 집합이 될 수도 있는 것이다. 필터와 다르게 래퍼는 연관성 점수를 개개의 자질에서 정의하지 않고, 선택된 자질 집합에서 정의한다. 그리고 그때의 연관성 점수는 통계치들을 사용하는 것이 아니라, 선택된 자질 집합들로 기계학습을 진행했을 때의 정확률을 사용한다. 결국 전체적인 과정은 특

정 자질을 기 선택된 자질 집합에 더했을 때 기계학습 결과의 정확률이 올라가면 그 자질을 선택하고, 정확률이 떨어지면 그 자질을 제외하는 방식으로 자질을 선택한다. 필터가 기계학습에 독립적이기 때문에 필터에서 높은 점수를 얻은 자질이라도 실제 기계학습에서의 성능은 나쁠 수 있는 반면, 래퍼는 기계학습에 의존적이기 때문에 래퍼에서 좋은 점수를 얻은 자질은 반드시 기계학습에 효과적이다. 이처럼 래퍼는 상당히 실용적인 방법이기에 일반적으로 필터보다 좋은 성능을 보이지만, 필터에 비해서 시간이 많이 걸린다는 단점이 있다.

2.2. 문서 분류와 특허 분류

문서 분류는 문서의 내용에 따라 문서를 미리 정의된 범주로 자동 분류하는 작업이다. 기존의 연구들은 이를 위해 지도식 기계학습 방법을 사용해 왔다. 기계학습 방법은 범주를 결정하는 규칙을 사람이 결정하는 것이 아니라 기계 스스로 학습한다는 점에서 효과적이며 특히 지도식 방법은 각 범주에 해당하는 문서들이 충분히 주어져 있을 때 적합하다. 일반적으로 문서분류에서 사용하는 자질은 그 문서에 등장하는 단어들이며, 이러한 지도식 기계학습을 통한 문서의 자동분류는 상당히 효과적인 방법임이 여러 연구를 통해 입증되어왔다.[1,2,3,9]

특허 분류는 문서 분류의 한 종류라고 볼 수 있으며, 현재까지는 일반적인 문서 분류와 비슷하게 지도식 기계학습 방법을 이용한 자동 분류가 연구되어 왔다. 하지만 특허 분류는 일반 문서 분류보다 좀 더 어려운 특징이 두 가지 있다. 우선 훈련집합과 평가집합이 방대할 뿐만 아니라 특허문서 자체의 길이도 일반 문서들보다 무척 길다. 훈련집합과 평가집합이 큰 것은 지도식 기계학습에 유리한 조건으로 생각될 수 있지만, 특허의 경우 그 양이 과하게 많기 때문에 자질이 너무 많아진다. 즉, 특정 특허문서의 분류에 관련이 없는 자질들도 너무 많이 학습에 포함되게 된다. 따라서 이것은 학습의 정확률을 떨어트릴 뿐 아니라 학습시간을 무척 길게 만든다. 두 번째로 분류 범주가 무척 다양하다. 특허는 모든 산업 분야에서 만들어 질 수 있기 때문에 다양한 특허가 존재할 뿐만 아니라 분류 자체가 무척 세부적으로 이루어져야 하기 때문에 분류 범주가 많다. 따라서 훈련집합이 많음에도 불구하고 일부 분류에는 자료 희귀 문제가 발생할 수 있다. 따라서 기계학습 방법을 통해 특허를 효과적으로 분류하기 위해서는 특허 문서 전체를 학습하지 않고 분류에 핵심이 되는 자질만을 선택해 학습하는 것이 필요하다.

3. 접 근 방 법

3.1. 필터 설계

문서분류에 사용되는 필터 방법에서는 일반적으로 연관성 점수로 TF, TFIDF등의 통계치를 사용한다. TF는 단어 출현빈도를 의미하며 단순히 자주 나타나는 단어는 중요한 단어일 것이라는 생각에 기초한 통계치 이다. TFIDF는 TF를 보완하기 위해서 단어의 출현빈도인 TF와 그 단어가 출현한 문서의 빈도인 DF의 역수를 곱해준 통계치 이다. TFIDF가 의미를 가지는 이유는 적은 문서에 등장하지만 해당 문서들에서는 자주 등장하는 단어가 핵심단어일 것이라는 추론에 근거한다. 많은 실험에서 TFIDF는 TF보다 좋은 결과를 보여왔다.[2,9] 본 연구에서는 TF, TFIDF와 더불어 새롭게 TFICF라는 새로운 통계치를 제안한다. TFICF는 TFIDF와 유사한 통계치로서 적은 분류 범주에 등장하는 단어에 높은 가중치를 준 것이다. 적은 분류 범주에 나타날수록 그 분류 범주를 결정 짓는데 핵심적인 단어일 가능성이 높기 때문이다.

일반적으로 필터 방법을 자질선택에 적용하는 과정은 다음과 같다. 우선 필터에 어떤 통계치를 사용할 것인지 결정하고 통계치의 임계값을 결정한 뒤, 임계값 이상의 모든 자질을 선택한다. 문제는 어떤 통계치를 사용하고 임계값은 몇으로 설정할 것인지를 어떻게 결정하느냐는 데에 있다. 본 논문에서는 이것을 래퍼를 이용해 결정하는 방법을 제안한다. 이를 위해 필터를 정적으로 고정시키지 않고, 통계치와 임계값 및 몇 가지 변수를 입력으로 주면 그에 해당하는 자질들을 가져올 수 있도록 동적으로 설계하였다.

3.2. 래퍼 설계

문서분류에 래퍼 방법을 그대로 적용하는 것에는 조금 문제가 있는데, 그것은 앞서 언급했듯이 문서분류에 사용되는 자질이 너무 많다는 점이다. 래퍼 방법은 자질들을 하나씩 선택하여 기계학습 결과에 따라 취하거나 버리는데, 수만 개에 이르는 자질이 존재하는 문서분류작업에 래퍼 방법을 적용하기 위해서는 수만 번의 학습과 평가가 필요하게 되는 것이다. 이를 해결하기 위해서 본 논문에서는 필터와 래퍼를 함께 사용하는 방법을 제안한다. 래퍼를 사용하되, 자질들을 하나씩 선택하는 것이 아니라, 필터를 변화시켜 가며 필터에서 한번 걸러진 자질들을 한꺼번에 선택하는 것이다. 그리고 그때의 기계학습 결과에 따라 취하거나 버린다.

필터를 결정하는 변수는 크게 4가지이다. 우선 사용할 수 있는 통계치는 3가지 이고(TF, TFIDF, TFICF), 임계값은 0부터 해당 통계치의 최대값까지의 실수를 선택할 수 있다. 또한 임계값은 절대값을 가지거나 평균으로부터의 상대값을 가질 수도 있으며, 실험에 사용된 특허문서의 구조는 크게 4부분으로 나눠지기 때문에 16가지의 조합 중 하나를 선택해 그곳에서만 자질을 추출할 수 있도록 하였다. 또한 이처럼 필터의 결정에 영향을 미치는 변수들이 서로 모두 독립적이라면, 각각의 변수들의 최적 값을 찾아 조합하면 최적의 해를 구할 수 있겠지만, 각 변수들은 독립적이 아니다. 예를 들어, TF를 필터의 통계치로 사용할 경우 TF는 평균도 낮고 편차도 작기 때문에, 임계값이 조금만 올라가도 성능이 크게 떨어지는 반면, TFIDF의 경우 평균도 높고 편차도 높아서 임계값이 올라가도 성능이 급격하게 하락하지는 않는다. 즉, 변수들이 서로 의존적이라는 것이다. 따라서 문제는 결국 다변수 함수의 최고점을 찾는 문제와 유사해 지게 된다.

우리는 이 다변수 함수의 최고점을 찾기 위해 시뮬레이티드 어닐링 방법을 사용하였다. 구체적인 방법은 매번 변수들의 값을 시뮬레이티드 어닐링 방법을 이용하여 결정하고, 결정된 변수에 해당하는 자질들을 추출(필터)한 후, 해당 자질들 만으로 훈련집합에서 기계학습을 하고, 기계학습을 통해 만들어낸 모델로 평가집합의 결과를 예측한 후, 평가를 통해 점수를 얻어 그것을 우리가 구하고자 하는 다변수 함수의 함수 값으로 삼았다.(래퍼) 결국 이 같은 과정을 통해 최대값을 가지는 변수에 해당하는 필터를 선택할 수 있었다.

4. 실험 및 평가

4.1. 평가 방법 및 실험 환경

기계 학습 모델을 통한 예측 결과는 상위 1000위까지 추출하였고, 평가방법은 TREC에서 사용하는 표준 평가방법인 MAP(Mean Average Precision)을 사용하였다. MAP은 재현율을 반영한 정확률로서 각각의 재현률 수준에서의 정확률을 모두 더한 뒤 평균을 낸 값이다. 따라서 정답이 예측결과의 상위에 있을수록 높은 값을 가지게 된다. 최종적으로 각각의 질의 문서에 해당하는 MAP값을 평균 내어 최종적인 평가점수로 사용했다.

실험 환경은 다음과 같다. 우선 래퍼에 사용한 기계 학습 방

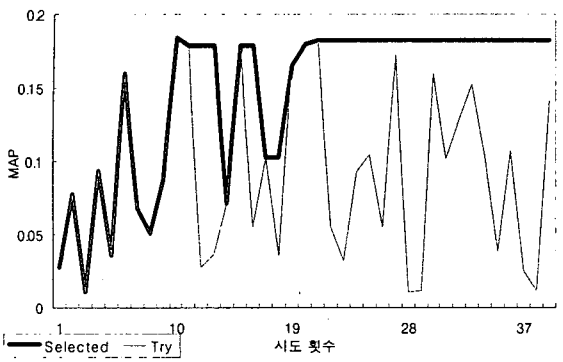

그림1. 순수한 시뮬레이티드 어닐링

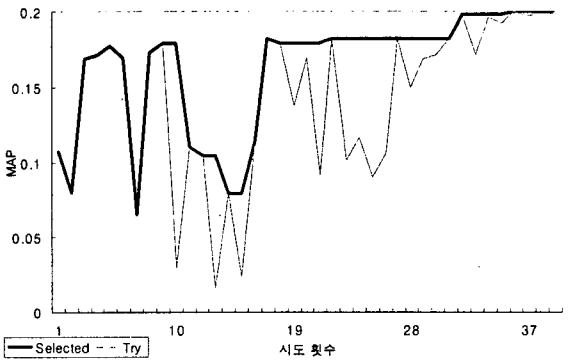

그림2. 휴리스틱을 적용한 시뮬레이티드 어닐링

법은 최대 엔트로피 모델이다. 훈련집합은 NTCIR-5 특허 분류 과제의 훈련집합 중 2344개의 문서를 균일한 분포로 추출하여 사용했으며, 평가집합은 마찬가지로 NTCIR-5 특허 분류 과제의 훈련집합 중 정답이 있는 450개의 문서를 추출하여 사용했다. 또한 모든 단어가 아닌 명사만을 추출 해 자질로 사용했다.

4.2. 실험 결과

<그림1>은 순수한 시뮬레이티드 어닐링 방법을 적용하였을 때의 결과이다. 결과를 통해 알 수 있는 것은 우선 시뮬레이티드 어닐링 방법의 특징에 의해서 MAP의 변화 양상이 크게 두 단계로 나누어 진다는 것이다. 첫 번째 단계는 수용 단계로 다양한 시도의 값이 좋든 나쁘든 대부분 그 값을 선택한다. 하지만 후반부에 접어들면서 양상은 두 번째 단계인 수렴 단계로 접어든다. 아직도 시도 자체는 다양하지만 알고리즘은 더 이상 모든 시도를 선택하지 않고 가능한 좋은 변화만을 선택한다. 첫 번째 실험의 가장 큰 특징은 각각의 시도가 완전히 무작위적이라는 점이다. 이런 완전 무작위 방법은 지역적인 최대값에 빠지는 문제를 막을 수는 있지만, 좀 더 좋은 값을 선택하기 위한 노력이 전혀 없어서 최대값으로의 수렴 속도가 늦어질 수 밖에 없다. 또한 가끔 너무 엉뚱한 값을 선택하기도 한다. 특히 후반부의 수렴단계에서는 완전 무작위적 방법을 통해서는 현재보다 좋은 값을 얻는 것이 힘들다.

따라서 두 번째 실험에서는 이를 보완하기 위한 몇 가지 휴리스틱한 방법들을 사용하였다. 우선 전혀 엉뚱한 값을 취하는 것을 막기 위해 선택할 수 있는 임계값의 범위를 어느 정도 제한하였다. 그리고 알고리즘이 수렴 단계에 접어들면 좀 더 빨리 최대 값에 수렴하게 하기 위해 두 가지 방법을 사용했는데, 엘리티즘과 기울기가 그것이다. 수렴 단계에 접어들면 우선 엘리티즘을 적용해 그 동안 가장 좋은 성능을 보였던 자질의 통계치와 임계값 기준을 선택할 확률을 크게 한다. 그리고 수렴 마지막 부분에서는 다른 변수들을 모두 고정 시키고 임계값 만을 변화시키는데, 기울기를 사용하여 이번의 시도가 성능을 좀더 높였으면 그만큼 정 방향으로 임계값을 증가시키고 성능을 낮췄을 경우 그만큼 역방향으로 임계값을 감소 시키는 그리디한 방법을 사용하였다. 이를 통해 어느 정도 선택된 자질집합의 최고 임계값에 다가가는 수렴속도를 조금 더 빠르게 만들 수 있었다.

<그림2>는 두 번째 실험의 결과를 보여주고 있다. 전체적으로 의도한 바대로 움직이는 양상을 관찰할 수 있었으며, 최종적으로 구한 최적 자질은 특허문서의 요약과 청구항에서만 자질을 선택하고, 통계치로 TFICF를 사용했으며, 임계값을 각 문서 집합의 TFICF 평균보다 -11.2만큼 낮게 잡았을 경우였다. 그리고 그 때의 MAP 값은 0.1997이었다.

5. 결 론

본 연구의 목적은 최고의 정확률을 가지는 자질 집합을 자동적으로 선택하는 것이다. 실험을 통해서 제안한 방법이 빠른 시간 안에 어느 정도 최고의 정확률을 가지는 자질 집합을 선택할 수 있다는 점을 확인할 수 있었다.

본 연구의 기여는 필터와 래퍼의 단점을 서로 보완하기 위하여 두 가지 방법을 상호 보완적으로 이용한 것이라 할 수 있다. 너무 많은 자질 집합을 가지는 지도식 학습에서 래퍼를 사용하는 것은 훈련시간 때문에 상당히 무리가 있고, 필터는 최적의 필터를 결정짓는 변수들을 정하는 것이 어렵다. 따라서 기존의 래퍼를 변형하여 최적의 자질이 아닌 최적의 필터를 자동적으로 찾아내는 래퍼를 설계하였고, 실험을 통해 이것이 좋은 성능을 보인다는 것을 증명하였다. 특히 분야에 관해 아무런 선지식이 없는 상태에서도 좋은 필터를 자동적으로 결정할 수 있다는 점이 중요하다.

현재 필터와 래퍼를 결합하여 이용하는 접근방법은 인공지능 분야에서 조금씩 이루어지고 있다.[1,8] 하지만 기존의 연구는 필터로 자질 집합의 일부를 선택한 후, 좀더 작은 자질 집합을 선택하는 데에 래퍼를 사용한 것으로서 본 연구와는 차이가 있다. 또한 필터와 래퍼를 조합한 자질 선택 방법을 문서 분류, 특히 특허 분류에 적용한 시도는 본 연구가 처음이라는 데에 의의가 있다.

6. 참고문헌

[1] Isabelle Guyon, Andre Elisseeff. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 2003

[2] Yiming Yang, and Jan O.Pedersen. A comparative study on Feature Selection in Text Categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997.

[3] C. Apte, F. Damerau, and S. Weiss. Towards language independent automated learning of text categorization models. In Proceedings of the 17th Annual ACM/SIGIR conference, 1994.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. John Wiley Sons, USA, 2nd edition, 2001.

[5] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer series in statistics. Springer, New York, 2001.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 2002.

[7] A. Rakotomamonjy. Variable selection using SVM-based criteria. JMLR, 3:1357-1370, 2003.

[8] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. JMLR, 3:1229-1243, 2003.

[9] Eui-Hong Sam Han, and George Karypis, Centroid-Based Document Classification: Analysis Experimental Results, Principles of Data Mining and Knowledge Discovery, 2000