자동 판례분류를 위한 기계학습기법

장균탁
고려대학교 컴퓨터정보통신대학원 미디어공학과
e-mail:gt75@korea.ac.kr

Machine Learning Technique for Automatic Precedent Categorization

Gyun-Tak Jang
Dept. of Media Engineering.
Graduate School of Computer and Information Technology,
Korea University.

요 약

판례 자동분류 시스템은 일반적인 문서 자동분류 시스템과 기본적인 동작방법은 동일하다. 본 논문에서는 노동법에 관련된 판례를 대상으로 지지벡터기계(SVM), 단일 의사결정나무, 복수 의사결정나무, 신경망 기법 등을 사용하여 문서의 자동 분류 실험을 수행하고, 판례분류에 가장 적합한 기계학습기법이 무엇인지를 실험해 보았다. 실험 결과 복수 의사결정나무가 93%로 가장 높은 정확도를 나타내었다.

1. 서론

사회에서는 다양한 원인으로 여러 가지 분쟁이 발생하게 된다. 이러한 분쟁을 조정하기 위한 법과, 법원에서 집행한 사건에 대한 판례를 검색할 수 있도록 다양한 법률관련 홈페이지가 운영 중이고, 법원에서 판결된 판례를 주제별로 분류하여 서비스하고 있다.

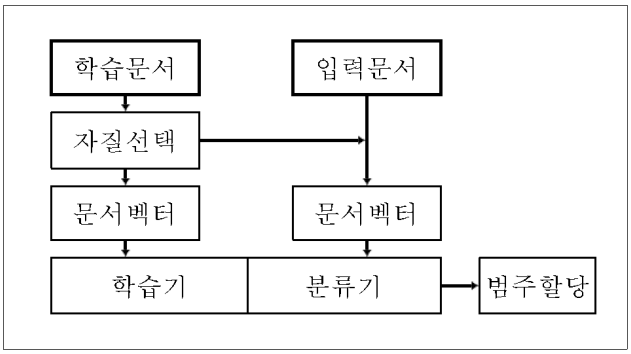
판례수집 및 분류업무는 법률전문가가 수작업으로 진행하여왔기 때문에 작업시간이 길어지고, 효율이 떨어진다. 작업효율을 높이기 위하여 자동 문서분류시스템을 판례에 적용하는 방법을 연구하게 되었다.

본 논문에서는 여러 가지 학습기법을 사용하여 판례를 자동으로 분류하는데 적합한 문서 분류 방법을 연구하여 본다.

2. 기계학습을 이용한 자동 문서분류

컴퓨터프로그램이 일련의 작업을 수행하면서 축척된 경험을 토대로 성능을 증가시켰다면, 그 프로그램은 학습을 했다고 하고, 이것을 기계학습이라고 한다.

일반적으로 기계학습을 기반으로 한 문서범주화 시스템은 전처리과정, 학습과정 및 분류과정으로 구성된다.



(그림 1) 자동 문서분류 시스템의 전체 구성도

(1) 전처리 과정

자동 문서분류시스템을 이용하기 위해서는 문서를 학습시스템에서 학습하기 쉽도록 변환하여야 한다. 이를 위해서 문서에서 특수문자 등을 제거하고, 웹문서인 경우에는 HTML문서에 포함되어 있는 태그와 특수 문자를 제거한다. 그리고, 형태소 분석기를 사용하기 위하여 문서의 내용을 문장 단위로 분할한다.

다음으로, 문서의 내용이나 특징을 잘 반영하는 단어를 추출해야 한다. 이를 내용어(content word)라 한다. 내용어를 추출하기 위해서는 형태소 분석기를 사용하여 문장을 각 형태소 별로 나누어 품사를 결정한다. 한국어에서는 개념을 설명하는데 쓰이는 명사가 가장 중요한 품사이다. 이렇게 추출된 내용어 중에 여러 문서에서 공통적으로 많이 나타나기 때문에 별다른 정보를 주지 못하는 불용어에 해당하는 내용어는 제거한다.

마지막으로, 불용어를 제외한 내용어들 중에서 범주화 구분에 유용하게 사용될 만한 내용어를 자질(feature)로 선택한다. 학습 문서에 나타나는 모든 내용어가 자질로 선택된다면 학습시간이 매우 오래 걸리게 되어 효율이 떨어지게 된다. 그러므로, 문서 범주화 성능의 저하 없이 자질의 수를 줄이기 위하여 차원축소 작업이 필요하다.

본 논문에서는 내용어중에 명사만을 추출하고, 추출된 명사 중에 학습문서집합에서 문서 내에 출현한 명사들의 문서 내 빈도(tf : term frequency)가 높은 명사 순으로 자질을 선택 하였다.

선택된 자질을 이용하여 학습에 사용될 문서를 색인화 하여 문서 벡터를 생성한다. 이것은 문서 전체에 나타난 각 자질의 출현 빈도를 이용하여 문서를 하나의 벡터로 표현하는 것으로, 본 논문에서는 문서 내 빈도(tf : term frequency)를 이용하여 문서벡터를 표현하였다.

(2) 학습과정

문서 분류 모델에는 기계 학습 분야에서 사용되는 여러 가지 알고리즘들이 사용된다. 본 논문에서는 주로 사용되는 3가지 알고리즘을 이용하여 실험을 진행하였다.

(가) 지지벡터기계 (SVM:Support Vector Machines)

지지벡터기계는 두개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik에 의해 소개된 학습기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리할 수 있는 결정면(decision surface)을 찾는 모델이다.

지지벡터기계는 직선으로 나눌 수 있는 문제에 사용되는 알고리즘이지만, 다차원의 부드러운 곡선을 이용하여 경계면을 설정하거나, 실제 데이터 벡터를 새로운 자질을 포함한 새로운 벡터 공간에 매핑하는 방법을 통해서 직선으로 나눌 수 없는 문제도 해결할 수 있다.

(나) 의사결정나무 (Decision Tree)

의사결정나무 기법은 개체의 속성(입력변수)과 부류로 구성된 데이터로부터 순환적 분할(recursive partinging)방식을 이용하여 나무를 구축하는 기법으로 구축되어진 나무는 나무의 가장 상단에 위치하는 뿌리마디(root node), 속성의 분리기준을 포함하는 내부마디(internal nodes), 마디와 마디를 이어주는 가지(link), 그리고 최종 분류를 의미하는 잎(leaver)으로 구성되며, 분류나 예측에 주로 사용된다.

의사결정나무기법은 데이터에 포함된 부류 값의 수에 관계없이 하나의 나무를 만드는 단일 의사결정나무(Single Decision Tree)와 데이터에 포함된 부류 값별로 의사결정나무를 구축하고 이들로부터 만들어지는 규칙들을 통합하는 복수 의사결정나무(Multi Decision Tree)로 나눌 수 있다.

의사결정나무에서는 데이터의 노이즈와 데이터의 과대맞춤 현상을 줄이기 위해 가지치기기법을 사용한다. 단일 의사결정나무 기법을 사용하여 모형은 구축하면 분류값들의 주요 속성들은 나무의 하단에 위치할 수 있다. 이러한 속성들은 가지치기 기법에 의해 제거될 수도 있다. 그러나 복수결정나무 기법을 사용하면 사례를 분류하는데 있어서 가장 중요한 속성을 우선적으로 고려하기 때문에 이들은 의사결정나무의 윗부분에 위치하게 되고 중요도가 떨어지는 속성들이 우선적으로 제거된다. 따라서 가지치기로 인한 주요 속성의 잘못된 제거를 줄일 수 있다.

(다) 신경망 (Neural Network)

신경망은 망(Network)의 성질을 가지고 있다. 즉, 가중치와 방향이 있는 에지(edge)로 구성된 그래프이다. 신경망은 은닉마디와 은닉계층을 몇 개로 하느냐에 따라 다양한 아키텍처를 구성할 수 있고, 특정 데이터에 대해 최적의 아키텍처가 무엇인지는 정해져 있지 않다. 본 논문에서는 이러한 신경망 알고리즘 중 가장 일반적인 방법인 멀티레이어 퍼셉트론(MLP:Multi-Layer Perceptrons)방법을 이용하였다.

3. 실험 및 분석

본 논문에서는 자동분류를 위해 지지벡터기계(SVM), 단일 의사결정나무, 복수 의사결정나무, 신경망 기법을 사용한다.

실험에 사용되는 판례는 (주)중앙경제에서 운영하는 홈페이지(www.elabor.co.kr)에서 28개의 분류로 서비스하고 있는 노동관련 판례 중 많이 사용되는 11개의 분류에 대하여 표1과 같이 분류별로 100개씩을 선택하여 총 1100개의 판례를 이용하였다.

<표 1> 실험에 사용된 분류

분류번호	분류	판례수
1	근로기준법 위반 관련	100
2	근로계약서, 취업규칙 관련	100
3	임금, 퇴직금 관련	100
4	근무시간, 휴일, 휴가 관련	100
5	해고, 징계 관련	100
6	산업재해보상보험 관련	100
7	업무상 재해 관련	100
8	노동조합 조직, 운영 관련	100
9	단체협약 관련	100
10	쟁의행위 관련	100
11	부당노동행위관련	100
계		1100

판례를 문서 벡터로 생성하기 위해 전체문서에서 가장 많이 쓰인 명사 100개를 자질로 선택하여 벡터 공간모델을 작성하였다. 단, 모든 판례에 공통으로 사용되는 피고, 원고, 주문, 청구취지, 이유, 결론 등의 단어는 불용어로 처리하여 자질에서 제외하였다.

문서 분류기는 상용되어 있는 toolkit들을 사용하였다. 지지벡터기계(SVM), 단일 의사결정나무, 복수 의사결정나무는 DTREG 4.5를 이용하였고, 신경망 기법은 SPSS Neural Connection 2.0 을 이용하였다.

각 분류기에 대한 실험결과는 정확도, 재현율, F-Measure로 계산하였고, 표2과 같다.

$$\text{정확도}(p) = \frac{\text{실제 범주가 } a \text{면서 } a\text{로 분류된 문서의 수}(pt)}{\text{범주가 } a\text{로 분류된 문서의 수}}$$

$$\text{재현율}(r) = \frac{\text{실제 범주가 } a \text{면서 } a\text{로 분류된 문서의 수}(rt)}{\text{실제 범주가 } a\text{인 문서의 수}}$$

$$F_\beta = \frac{(\beta^2+1)\cdot \text{정확도} \cdot \text{재현율}}{\beta^2 \cdot \text{정확도} + \text{재현율}} \quad (\text{단, 이 실험에서 } \beta = 1)$$

<표 2> 판례 자동분류 실험결과

분류기 / 분류번호	지지벡터 기계(SVM)			단일 의사 결정나무			복수 의사 결정나무			신경망 (MLP)		
	p	r	F	p	r	F	p	r	F	p	r	F
1	0.878	0.720	0.791	0.800	0.480	0.600	0.935	0.860	0.896	0.773	0.680	0.724
2	0.829	0.680	0.747	0.493	0.740	0.592	0.967	0.870	0.916	0.660	0.700	0.679
3	0.610	0.890	0.724	0.574	0.700	0.631	0.800	0.960	0.873	0.646	0.620	0.633
4	0.896	0.860	0.878	0.803	0.530	0.639	0.938	0.900	0.919	0.759	0.820	0.788
5	0.796	0.860	0.827	0.629	0.440	0.518	0.979	0.940	0.959	0.700	0.700	0.700
6	0.837	0.820	0.828	0.783	0.720	0.750	1.000	0.940	0.969	0.867	0.780	0.821
7	0.870	0.870	0.870	0.646	0.840	0.730	0.962	1.000	0.981	0.733	0.850	0.787
8	0.867	0.780	0.821	0.764	0.550	0.640	0.957	0.880	0.917	0.787	0.740	0.763
9	0.891	0.820	0.854	0.679	0.760	0.717	0.960	0.960	0.960	0.955	0.840	0.894
10	0.920	0.920	0.920	0.611	0.660	0.635	0.933	0.970	0.951	0.843	0.860	0.851
11	0.830	0.880	0.854	0.610	0.720	0.660	0.875	0.980	0.925	0.741	0.830	0.783
평균	0.839	0.827	0.829	0.672	0.649	0.647	0.937	0.933	0.933	0.769	0.765	0.766

실험결과 복수 의사결정나무가 가장 높은 정확도를 보여주었고, 지지벡터기계, 신경망기법, 단일 의사결정나무 순의 정확도를 보여주었다.

4. 결론

판례를 자동으로 분류하기 위하여 본 논문에서는 자동분류를 위해 지지벡터기계(SVM), 단일 의사결정나무, 복수 의사결정나무, 신경망 기법을 사용하여 실험을 하였다. 실험 결과 지지벡터기계(SVM)와 복수 의사결정나무를 이용한 분류에서 각각 82%이상, 93%이상의 높은 정확도를 보여주었다. 자질선택과 문서벡터 생성시의 가중치를 조절한다면 정확도를 더욱 높일수 있을 것이다.

판례문서의 경우는 다른 문서에 비해 사용하는 단어도 다르고, 문서사이의 유사도가 높기 때문에 이 실험결과가 의미가 있다고 판단된다.

향후, 다른 분야의 판례에도 적용해 보고, 색인방법을 변경하고, 다른 분류기법을 이용하여 분류의 정확도를 높이는 방법을 더욱 개발해야 할 필요가 있다.

참고문헌

[1] 고영중, 서정연, “문서관리를 위한 자동문서범주화에 대한 이론 및 기법”, 정보관리연구 vol.33, no.2, pp.19-32, 2002.

[2] 김상범, “범주간의 상호관계를 고려한 자동 문서범주화의 개선”, 고려대학교 컴퓨터학과 전산학석사 학위논문, 1999.

[3] 남상엽, “카테고리간의 상하위 관계를 고려한 문서자동분류 시스템”, 창원대학교 컴퓨터공학과 석사 학위논문, 2001.

[4] 권용진, 안준선 역, “정보검색 알고리즘”, 미래컴, 2003.

[5] www.dtreg.com, “DTREG : Predictive Modeling Software”, 2007.

[6] 신은주, 장남식, “데이터마이닝 기법 비교 연구 : 단일 및 복수 의사결정나무”, 한국경영정보학회 춘계학술대회논문집, pp.361-369, 1999.

[7] 김시환, 권영식, “데이터마이닝을 이용한 인터넷 쇼핑몰 고객세분화”, SPSS 사용자 사례 논문, pp.41-69, 2000.