정보관리연구, vol.39, no.4 2008, pp.47-66

기계학습을 이용한 문서 자동분류에 관한 연구

A Study on the Documents's Automatic Classification Using Machine Learning

김 성 희* · 엄 재 은**
Seong-Hee Kim · Jae-Eun Eom

차 례

1. 서 론
2. 이론적 배경
3. 연구설계
4. 실험수행
5. 결 론
• 참고문헌

초 록

본 논문에서는 수작업 분류 시 초래하는 여러 가지 한계점을 극복하고, 이용자에게 보다 빠르고 정확한 분류 서비스를 제공하기 위해 4개의 다양한 기계학습 알고리즘을 적용하였다. 연구대상으로는 MeSH의 8개의 주제별 범주로 각각 100개의 문헌 타이틀을 선정하였으며, 4개의 기계학습 알고리즘으로 실험을 수행하였다. 그 결과 신경망 기법과 C5.0 기법을 병행하여 사용했을 경우 단일 기법을 사용했을 경우보다 2.5%, 3.75%가 상승하여 분류 효율이 83.75%로 측정되었다. 이 수치는 4개의 분류 실험 결과 중 가장 높은 정확률을 나타내었다. 따라서 신경망 기법과 C5.0 기법의 장점을 이용하여 분류 서비스를 실행하면 단일 기법을 사용한 경우보다 높은 정확률을 나타낼 수 있을 것이라 기대된다.

키 워 드

신경망, 의사결정나무, 문서 자동분류, 기계학습

* 중앙대학교 문헌정보학과 부교수
(Associate Professor, Dept. of Library & Information Science, Chung-Ang University, seonghee@cau.ac.kr)
** 중앙대학교 대학원 문헌정보학과
(Graduate Student, Chung-Ang University, east81@nate.com)
• 논문접수일자 : 2008년 8월 25일
• 게재확정일자 : 2008년 9월 19일

ABSTRACT

This study introduced the machine learning algorithms to overcome the many different limitations involved with manual classification and to provide the users with faster and more accurate classification service. The experiments objects of the study were consisted of 100 literature titles for each of the eight subject categories in MeSH. The algorithms used to the experiments included Neural network, C5.0, CHAID and KNN. As results, the combination of the neural network and C5.0 technique recorded classification accuracy of 83.75%, which was 2.5% and 3.75% higher than that of the neural network alone and C5.0 alone, respectively. The number represented the highest accuracy rates among the four classification experiments. Thus the use of the neural network and C5.0 technique together will result in higher accuracy rates than the techniques individually.

KEYWORDS

Neural Network, Maching Learning, KNN, Decision Tree, Automatic Document Classification

1. 서론

문서분류란 문서의 내용을 작업자가 읽고 문서를 미리 정의한 범주로 분류하는 작업이다. 일반적으로 분류라고 하면 정해진 분류체계 하에서 분류하고자 하는 각 문헌들을 가장 적합한 카테고리로 배정하여 집단화하는 작업을 의미한다. 이전에는 문서분류를 수작업으로 수행하였으나 이 경우 사람의 노력, 시간, 비용 면에서 어려움을 초래하였다. 이러한 어려움은 기존의 문서분류를 자동분류시스템으로 교체하거나 보조시스템을 활용하면 여러 면에서 작업의 효율성을 높일 수 있다. 문서자동분류는 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시킬 수 있다.

본 연구에서는 자동분류시스템 중에서 주어진 데이터를 분석하여 내재되어 있는 중요한 패턴이나 규칙성을 제시하는 기계학습의 분류 알고리즘을 이용하여 문서자동 시스템의 효율성을 알아보고자 한다. 기계학습 방법에 의한 문서분류는 작업 비용을 최소화시킬 수 있으며 객관적인 기준으로 문서를 분류함으로 문서분류 문제에 대한 효과적인 해결책을 제시할 수 있다(박성태, 태윤식 2007). 즉, 본 연구를 통해 수작업 분류 시 초래하는 여러 가지 한계점을 극복하고, 기계학습 알고리즘을 도입해 이용자들에게 보다 빠르고 정확한 분류 서비스를 제공할 수 있으며 이용자들이 필요로 하는 자료를 방대한 양의 데이터 안에서

선택적으로 추출하게 함으로써 보다 효율적인 분류 서비스를 수행할 수 있을 것이라 기대한다. 본 논문의 연구대상으로는 MeSH의 8개의 주제별 범주로 각각 100개의 문헌 타이틀을 선정하였으며, 기계학습 개념을 이용하여 자동적으로 문서를 분류하는 방법을 제시 하였다. 제시된 방법으로는 Neural Network, C5.0, CHAID, K nearest neighbor Classifier의 4개의 기계학습 알고리즘으로 분류 실험을 수행하였다. 연구의 제한점으로는 실험에 필요한 데이터를 미국 국립의학도서관에서 구하여 실험하는 한정된 데이터 사이즈이다.

2. 이론적 배경

2.1 신경망(Neural Network)

신경망 모형은 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 흉내 내 자신이 가진 데이터로부터 반복적인 학습 과정을 거쳐 패턴을 찾아내고 이를 일반화하는 모형이다. 특히 향후를 예측하고자 하는 문제에 있어 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형의 하나이다. 주로 감독학습에 적용되며 결과변수에 대한 예측이나 분류를 목적으로 감춰진 패턴을 찾고 이를 일반화하는 데 이용한다.

신경망 기법의 기본적인 알고리즘으로는 역전파 알고리즘(Back Propagation)이 있으며, 여기에는 입력과 출력 노드 및 은닉 노드로 구성되어 있다. 입력층에 값을 주면 이 신호가 각 뉴런에서 변환되어 중간층에 전달되고, 출력층에서 신호를 출력하게 되는데 이렇게 얻어진 출력값과 목표값을 비교하여 오차를 줄여 나가게 된다. 실제 출력과 정확한 값에 해당하는 기대 출력값을 비교하고, 여기에 차이가 있을 때는 각 노드와 네트워크의 시냅스에 대한 가중치를 조정한다. 이러한 과정은 응답결과가 어느 정도 정확할 때까지 반복 시행되며 신경망의 구조가 안정화되면 학습 단계는 종료가 되어 신경망 모형을 산출하게 된다.

2.2 의사결정나무(Decision Tree)

의사결정나무는 의사결정규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 분석과정이 나무구조에 의해 표현되며 분류나 예측을 목적으로 하는 다른 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있는 장점이 있다(최종후 외 2000, 17). 즉 질문의 과정이라고 볼 수 있는데 첫 번째 질문이 다음에 가야할 경로를 결정하게 되며, 질문이 잘 선택될수록 짧은 시간에 들어온 레코드를 효율적으로 분류할 수 있다. 처음에 선택할 분류기준에는 여러 알고리즘이 있으며 입력된 레코드가 끝마디에 갈 때까지 이러한 과정은 반복된다.

의사결정나무의 생성 단계는 첫째, 의사결

정나무의 형성(Growing), 둘째, 가지치기(Remove branch), 셋째, 타당성평가(Assessment), 넷째, 해석(Interpretation) 및 예측(Prediction)등의 과정을 거쳐서 수행된다. 각 단계별로 살펴보면 첫째, 의사결정나무의 형성이다. 여기에는 분석의 목적과 자료구조에 따라서 적절한 분리기준과 정지 규칙을 지정하여 의사결정나무를 얻도록 한다. 둘째, 가지치기로 분류 오류를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거한다. 셋째, 이렇게 제거한 가지의 타당성을 평가한다. 이익도표나 위험도표 또는 검증용 자료에 의한 교차타당성 등을 이용해 의사결정나무를 평가하도록 한다. 넷째, 평가한 의사결정나무를 해석하고 예측모형을 설정한다.

2.3 K nearest neighbor Classifier

KNN(K nearest neighbor Classifier)은 Yang(1994)에 의해서 제안되었다. 이 알고리즘의 특징은 아주 간단하다는 것과 학습단계에서 최소한의 처리 작업을 한다는 것이다. 문서분류 관점에서 KNN의 정확도는 매우 우수하나, 분류 단계에서 실행속도는 매우 느리다는 단점을 가지고 있다.

KNN은 학습 단계에서는 학습 문서에 대한 벡터화정도만 한다. 따라서 KNN이 어떻게 동작하는가는 분류단계만의 설명으로 충분하다. KNN은 분류집합, 학습문서 집합, 시험 문서 및 상수 K를 입력받는다. 이러한 입력으로부터 모든 학습 문서에 대하여 시험문서와의 유사성을 계산하여 유사성이 가장 큰 K개의 학습 문서를 선택하고, 선택된 학습 문서가 속하는 분류들에 대하여 등급을 정하여 이들을 출력한다(이재문 2003).

2.4 선행연구

다양한 기계학습 알고리즘을 이용한 자동문서분류에 관한 연구가 수행되어 왔다.

Yang(1999)은 다양한 12 종류의 문서 범주화 기법의 성능을 직, 간접적으로 비교하였다. 비교 결과 전반적으로 KNN, LLSF, 신경망이 가장 높은 성능을 보여주는 것으로 실험결과 나타났으며, Bayes 기법을 제외한 다른 기계학습 알고리즘들도 비교적 높은 성능을 보여주는 것으로 나타났다. Yang (1999)의 또 다른 연구에서는 다섯 개의 분류기의 성능을 비교하였다. 비교 대상 분류기는 SVM(Support Vector Machine), KNN, 신경망, the Linear Least-squares Fit(LLSF) mapping, Native Bayes 분류기이다. 이 연구의 실험 결과 범주가 충분히 일반적이면서 범주 당 적합한 학습 문서의 수가 10개 이하로 적을 경우에 SVM, KNN 및 LLSF가 신경망이나 Native Bayes보다 훨씬 높은 성능을 보여주는 것으로 나타났다.

Ruiz and Scrinivasan(1999)은 의학 분야 문헌(MEDLINE)의 MeSH의 주제명을 자동부여하기 위한 목적으로 로치오 알고리즘과 역

전파 알고리즘을 사용하였다. MeSH 색인어를 119개를 선정하고 세 가지 방법(로치오 알고리즘, 신경망, 계층적 신경망)을 통하여 실험하였다. 실험결과 계층적 신경망의 성능이 가장 좋은 결과를 보였다.

Chang(2000)은 생명정보학(Bioinformatics) 분야의 논문에 MeSH 색인어를 부여하기 위해 Native Bayes 알고리즘과 KNN을 계층적, 비 계층적 두가지 방법으로 나누어 적용하였다. 그 결과 KNN보다는 Bayes가 더 좋은 성능을 보였고, 계층적 방법보다는 비 계층적 방법이 나은 성능을 보여주었다.

이영숙, 정영미(2001)는 계층적 분류체계인 DDC 분류표에 따라 문서를 자동으로 분류하기 위해 HiCat이라는 자동분류 알고리즘을 제안하였다. HiCat 알고리즘은 주제어 지식베이스와 학습 테이블을 동시에 이용하고, 각 계층에 속한 주제 범주들을 대상으로 주제 적합성 가중치 공식을 적용하여 범주를 결정해가는 방식을 취한다. 이 실험에서 Native Bayes, KNN과 이를 변형한 NB_hier, KNN_hier와의 성능 비교 및 평가를 하였다.

노영희(2001)는 기계학습을 기반으로 한 인터넷 학술문서의 자동분류에 관하여 연구하였다. 연구방법으로 KNN분류기를 이용하여 범주화 방법에 대한 성능 실험을 하였다. KNN분류기는 학습문서집단에서 검증문서와 가장 유사한 K개의 학습문서를 찾아야 하는데, 이에 가장 적합한 K값은 얼마인지를 실험을 통하여 검증하여 보고자 하였다. 그 결과 자질 축소비율이나 K값의 변화에 관계없이 문서가 많이 분류되었으며 재현율과 정확률은 90% 이상의 높은 성능을 보여주고 있는 것으로 나타났다.

김진화(2004)는 인공지능을 이용하여 웹 문서를 자동 분류하였다. 연결빈도 행렬을 이용한 문서지문기법을 자동분류 방법으로 제안하였다. 연결빈도 행렬과 4가지 학습 알고리즘(Keyword matching, Native Bayesian, 의사결정나무, 신경망을 이용하여 분류 정확률을 비교, 분석하였다. 그 결과 분류 정확도는 두 번째로 높게 나타났다.

황성하(2005) 외 3인은 인터넷 문서의 자동분류 서비스 시스템에 관하여 연구하였다. 인터넷 정보검색은 에이전트와 분류, 변환 등의 가공 기술에서 더욱 두드러지게 발달하고 있으며, 시스템의 자동화를 통한 편의성을 제공한다면 더욱 효과적인 정보관리가 이루어질 것이다. 인터넷 정보의 수집에서 자동분류, 검색서비스까지를 하나의 시스템으로 처리할 수 있는 인터넷 문서 자동분류 서비스 시스템을 소개하였다.

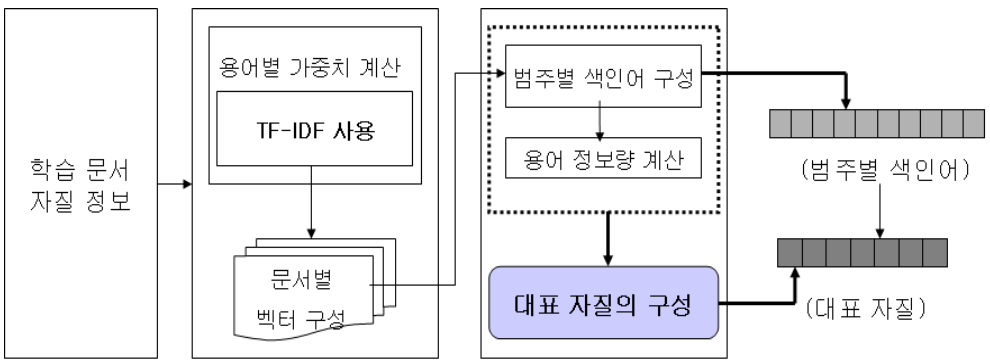
김판준, 이재윤(2007)은 문헌 간 유사도를 자질로 사용하는 분류기에서 미 분류 문헌을 학습에 활용하여 분류 성능을 높이는 방안을 모색하였다. 미 분류 문헌을 활용하는 준 지도 학습 알고리즘은 대부분 수작업으로 분류된 문헌을 학습데이터로 삼아서 분류하는 첫 번째 단계와, 수작업으로 분류된 문헌과 자동으로 분류된 문헌을 모두 학습 데이터로 삼아서 분류기를 학습시키는 두 번째 단계로 구성하

여 연구하였다. 그 결과 미 분류 문헌을 활용하는 것이 문헌 유사도 자질 표현 방식을 채택한 자동분류 시스템을 도입할 때 분류의 효율성을 높일 수 있는 대안이라고 제시하였다.

3. 연구설계

기계학습에 의한 문서분류는 학습하는 훈련 단계와 분류하는 시험 단계로 구분한다. 학습 단계는 미리 정해진 분류 집합과 각 분류별로 전문가에 의해서 정확히 분류된 학습 문서 집합을 입력으로 받아 학습하는 과정이다. 이 단계는 시험 단계에서 최적의 성능을 얻을 수 있도록 입력 데이터를 가공한다. 분류 단계는 이러한 가공된 데이터와 새로운 문서를 입력받아 입력된 문서가 어느 분류에 속하는지를 판단하는 단계이다. 따라서 학습 문서는 대부분의 경우 한번 실행되고, 분류 단계는 새로운 문서가 발생할 때마다 실행된다.

본 연구에서는 범주별 문서집합을 생성하기 위해 미국 국립의학도서관 MeSH의 8개의 주제별 범주를 선정하였고, 각 범주별로 100건의 데이터씩, 총 800건의 데이터를 추출하였다. 800건의 데이터는 해당 범주의 문서 타이틀을 그 대상으로 하였다. 원래 MeSH는 16개의 주제별 범주를 가지고 있으나, 이 중 8개는 의학 분야와 직접적인 관련이 없다고 판단하여 제외하였다. MeSH는 [Anatomy], [Organisms], [Diseases], [Chemicals and Drugs], [Analytical, Diagnostic and Therapeutic Techniques and Equipment], [Psychiatry and Psychology], [Biological Sciences], [Natural Sciences], [Anthropology, Education, Sociology and Social Phenomena], [Technology, Industry, Agriculture], [Humanities], [Information Science], [Named Groups], [Health Care], [Publication Characteristics], [Geographicals]의 16개 범주(2008)로 이루어져 있



〈그림 1〉 자질 선정 및 색인어 구성 흐름도

다. 본 논문은 16개의 범주 중에 [Anatomy], [Organisms], [Diseases], [Chemicals and Drugs], [Analytical, Diagnostic and Therapeutic Techniques and Equipment], [Psychiatry and Psychology], [Biological Sciences], [Natural Sciences]의 8개 범주만을 선택하였으며 각 범주별 표시는 1에서 8까지 순차적으로 숫자로 표현하였다.

범주별 문서집합의 자질추출을 위해 사용된 스테밍 방법으로는 현재 널리 사용되고 있는 포터 스테밍 알고리즘 (Porter Stemming Algorithm)을 사용하였다(〈그림 1〉 참조).

전 처리과정을 마친 전체 용어를 이용하여 전체 word 벡터를 구성하고, 범주별 대표 색인어를 산정한 후 입력받은 학습 문서의 자질 정보에서 용어별 가중치를 TF-IDF를 이용하여 계산하였다. 이렇게 산정된 범주별 색인어 집합에서 용어 정보량을 계산하여 대표 자질 360개가 추출되었다. 〈표 1〉은 범주 6과 범주 7의 상위 20개 자질의 자질을 나타내고 있다.

기계학습에 의한 문서분류의 학습과정은 전문가에 의해 정해진 범주로 분류된 학습 문서 집합을 입력받아 학습하는 과정이다. 총 800건의 문서에 대해 640건의 문서를 학습과정에

〈표 1〉 [Biological Sciences]와 [Psychiatry and Psychology]의 상위 20개 자질

No.	Biological Sciences		Psychiatry and Psychology	
	Keyword	Weight	Psycholog	7.456
1	Pregnanc	5.6901	Psychiatri	5.28012
2	Chiropract	4.17816	Cognit	4.6471
3	Chromosom	4.6954	Antipsychot	3.75517
4	Acupunctur	3.26599	Somatoform	3.66675
5	Phagocytosi	2.7385	Anorexia	3.4678
6	Synaptonem	2.67402	Dementia	3.4678
7	Centromer	2.60451	Nervosa	3.34394
8	Permeabl	2.52893	Agoraphobia	2.93534
9	Receptor	2.52893	Psychopharmacolog	2.67962
10	Estrous	2.44591	Creutzfeldtjakob	2.63932
11	Insemin	2.44591	Neurocirculatori	2.63932
12	Pigment	2.44591	Schizophren	2.40331
13	Telomer	2.44591	Parasomnia	2.34672
14	Ejacul	2.35344	Asthenia	2.21939
15	Embalm	2.35344	Comorbid	2.21939
16	Pharmacoepidemiolog	1.76594	Delirium	2.21939
17	Ethnopharmacolog	1.69168	Patholog	2.21939
18	Neuroanatomi	1.57428	Alzheim	2.19002
19	Techyphylaxi	1.57428	Enuresi	2.14652
20	Erythrocyt	1.50414	Mental	2.06538

〈표 3〉 800개의 학습문서와 360개의 자질을 이용한 학습 실험

학습 실험 번호	학습 네트워크의 구성			training step	error(%)
	입력 유닛	은닉 유닛	출력 유닛		
학습 실험 1	360	20	8	300	6.92
학습 실험 2	360	35	8	300	7.13
학습 실험 3	360	40	8	400	7.13
학습 실험 4	360	43	8	430	5.72
학습 실험 5	360	50	8	380	6.05
학습 실험 6	360	55	8	550	4.82
학습 실험 7	360	65	8	550	4.69
학습 실험 8	360	70	8	580	4.73

사용하였으며 160건은 실험 데이터로 사용하였다. 이렇게 선정된 데이터를 이용해서 서로 다른 기계학습 알고리즘을 적용하여 분류하는 실험을 수행하였다.

4. 실험수행

4.1 신경망 기법을 이용한 문서분류

신경망 기법은 인간의 두뇌에 있는 대규모의 뉴런들의 상호 연결되어 있는 구조를 모델링한 것으로 역 전파 알고리즘에 기본을 두고 있다. 일반적으로 역 전파 알고리즘은 매개 변수와 은닉 유닛의 수를 조정하여 학습 실험을 함으로 최소 에러값을 갖는 네트워크 가중치를 구하여 분류한다.

본 논문에서는 입, 출력의 노드수는 고정되어 있으므로 문제가 될 수 없지만 은닉층의 노드수는 문제가 되고 있다. 왜냐하면 은닉층의 노드 수에 따라 분류 결과가 달라질 수 있기 때문이다. 이것은 시스템 성능에 큰 영향을 미

〈표 2〉 모멘트(Alpha)와 학습율(Eta)의 정확률 비교

훈련자료			
Alpha	정확율	Eta	정확율
0.9	79.85%	0.9	83.19%
0.5	81.16%	0.5	80.77%
0.1	78.57%	0.1	82.54%

훈련자료		
Alpha	Eta	정확율
0.9	0.9	83.19%
0.9	0.5	83.77%
0.9	0.1	82.54%
0.5	0.9	81.23%
0.5	0.5	79.37%
0.5	0.1	80.95%
0.1	0.9	77.47%
0.1	0.5	81.89%
0.1	0.1	81.82%

치지만 아직까지 은닉층의 크기를 결정하는 것과 관련해서는 통일된 방법이 없다. 따라서 본 연구에서는 어떤 값이 신경망에 응용될 때 가장 이상적인지 경험적인 방법에 의하여 실험하였다. 여기서 경험적이라고 하는 것은 다양한 값을 적용해 봄으로 어떤 값이 가장 적절한지를 경험에 의해 결정하는 것이다. 신경망 기법은 주로 학습율과 모멘트가 사용되므로 이 수치를 어떻게 입력하느냐에 따라 모형 적합 시간은 영향을 받는다. 그렇기 때문에 분석하고자 하는 자료에 적당한 수를 선정하여 오차가 최소값이면서 학습과정이 빠르게 수렴할 수 있게 학습하도록 하는 것은 매우 중요한 문제이다(김천식, 홍유식 2006). 따라서 본 논문에서는 그 상관관계를 밝혀내기 위해 15가지의 분류 실험을 수행하였다. 〈표 2〉는 모멘텀과 학습율을 0.9, 0.5, 0.1로 실험해 본 결과이다.

그 결과 모멘텀 값이 0.9이며 학습률 값이 0.5일때 분류 효율이 83.77%로 가장 높았고 다음으로 모멘텀 값과 학습률 값이 둘 다 0.9일 때, 83.19%라는 효율성을 나타냈다. 따라서 본 논문에서는 모멘텀과 학습율을 0.9, 0.5로 지정하여 수행하도록 하였다. 이러한 결과는 기존의 여러 연구(Cherkassky and Vassilas 1989; Gersho and Reiter 1999) 결과와 유사하게 나타났다. 김성희(2000)는 신경망 시스템 설계 시 성능에 영향을 미치는 변수들을 경험적 방법을 통해 검토하였다. 실험에서는 학습율이 0.5, 모멘텀이 0.9이거나 학습율이 0.9, 모멘텀이 0.5 일 때 가장 분류 속도가 빠르게 나타났다. 따라서 학습시간을 최소화시키고 네트워크를 성공적으로 학습시키기 위해서는 학습율 및 모멘텀의 값을 시스템 설계에 고려해야 할 것으로 보인다.

〈표 3〉은 800개의 학습문서에 대해 360개

Correct	130	81.25%
Wrong	30	18.75%
Total	160	

- 대상 : test data(160건)
- 정확도 : 81.25%
- 입력 레이어 : 360개의 뉴런
- 출력 레이어 : 8개의 범주

〈표 4〉 신경망 기법의 분류 결과

범주명	test문서 수	적중한 문서 수	Precision(%)
Anatomy	20	15	75
Organisms	20	15	75
Diseases	20	16	80
Chemicals & Drugs	20	17	85
Analytical, Diagnostic and Therapeutic Techniques and Equipment	20	18	90
Psychiatry and Psychology	20	18	90
Biological Sciences	20	16	80
Natural Sciences	20	15	75
총 테스트 문서 수	**160**	**130**	

의 자질을 이용하여 자동분류를 수행한 결과이다. 그 결과 가장 작은 최소 에러값을 출력한 실험은 학습 실험 7이다.

학습 실험 7은 학습문서가 640개이고, 입력유닛이 360개, 출력 유닛이 8개인 역 전파 네트워크에서 은닉 유닛의 개수가 65개일 때 에러값이 4.69%로 최소 에러값을 나타냈다. 종합적인 실험 결과 본 논문에서는 train cycle은 500+이며 hidden layer를 사용하였으며, hidden node는 65개, 학습율은 0.5로, 모멘트는 0.9를 산정하였다.

160건의 실험 데이터에 관한 분류 결과 중 성공적으로 이루어진 데이터는 총 160건 중 130건으로 81.25%에 해당되며, 18.75%인 30건은 성공적으로 이루어지지 못했다. 범주당 정확률을 살펴보면 〈표 4〉와 같다.

〈표 4〉에서 나타난 바와 같이 [Analytical, Diagnostic and Therapeutic Techniques and Equipment]와 [Psychiatry and psychology] 범주가 다른 범주에 비해 정확률이 높은 것을 알 수 있다. 이는 비교적 자질의 특성이 뚜렷하기 때문이며 학습이 제대로 이루어졌다는 것을 의미한다. 반면에 그렇지 않는 [Natural Sciences]나 [Organisms]는 분류 정확률이 좋지 않았다. 특히 [Natural Science]의 경우, 두 가지 측면으로 그 원인을 살펴볼 수 있다.

첫 번째로 [Natural Sciences]의 범주에 속하는 문서이긴 하지만 그 자질은 다른 범주에 공통적으로 속하는 경우가 많았다. 예를 들면 neurophysiologi는 [Natural Sciences]의 대표 자질이지만 [Anatomy] 범주와 [Organisms], [Diseases]에 포함되어 있는 자질이다. 마찬가지로 neuroscience 또한 [Anatomy], [Psychiatry and Psychology] 범주에 포함되어 있었다. 이밖에도 여러 공통적인 자질의 존재로 [Natural Sciences]만의 특성화된 색인어가 다른 범주에 비해 효과적으로 추출되지 못했고 분류 효율성의 저하라는 결과를 가져온 것이라고 추측할 수 있다.

두 번째로, 학습문서와 실험문서의 자질 집합의 가중치 차이가 있었다. 예를 들어 [Natural Science]에 관련된 문서이지만 [Natural Science]에 분류되지 않는 문서의 자질을 살펴본 결과, star, solar, metric 등에 관한 내용을 포함하고 있었다. 초기 800건의 학습 문서에는 [Natural Science]에 대한 특성화된 자질이 Amplifi geographi, Oceanographi, Calibr 등에 관한 내용이 많았고 star, solar, metric에 관한 학습 문서는 거의 다루지 않았다. 따라서 자질 추출과정에서 이 단어들의 중요순위는 낮게 나타났으며 학습 데이터의 네트워크 가중치를 생성할 때도 마찬가지로 적용되었다. 하지만 실험 데이터에서는 이러한 자질의 내용을 가진 문서들이 있었으며, 이 부분에 대한 학습은 충분히 이루어지지 않았기 때문에 실험 데이터의 분류작업은 효율적으로 이루어지지 못한 것으로 보인다. 즉 학습 데이터와 실험 데이터의 가중치의 차이 때문에 [Natural Sciences]의 분류 정확률은 저하되

었다고 추측할 수 있다.

[Natural Sciences]의 오 분류는 자질 집합이 신경망 기법의 학습 네트워크 수행 시 학습이 제대로 이루어지지 않았음을 나타낸다. 즉 다양한 학습 문서를 선별하여 자질 집합을 구축하고 학습과정을 수행하는 것이 분류의 성능을 결정한다는 것을 알 수 있다.

4.2 의사결정나무를 이용한 문서분류

4.2.1 C5.0을 이용한 문서분류

C5.0 모델은 정보 획득값이 최대값을 가지는 입력 필드를 분류 필드로 하여 의사결정나무 모델의 알고리즘 중 가장 정확도가 높은 것으로 알려져 있으나 과잉적용의 문제가 있어 가지치기의 설정이 중요한 모델링 방법이다.

가지치기 방법에는 2가지가 있다. 하나는 일정한 임계치를 적용하여 그 이상인 경우에만 새 가지가 나오도록 허용한다. 다른 하나는 나무를 충분히 자라게 한 후 중요도가 비교적 작은 가지 분리들을 취소함으로 나무를 규모 감축시키도록 한다. 이 경우 계산시간이 늘어나지만 가지의 중요도를 확인한 후에 취사선택하기 때문에 그만큼의 가치가 있다. C5.0은 전신인 C4.5로부터 후자의 방법을 취하고 있다.

본 논문에서는 학습 데이터에 대한 출력유형은 Rule set으로 지정하였다. 또한 교차타당성을 확인하기 위해 접기 횟수를 10으로 설정하였으며, boosting 사용 역시 시도횟수를 10으로 제한하였다. 증폭이나 교차타당성의 횟수는 10에서 50까지 순차적으로 입력하여 실험하였으나 분류 효율성에 그다지 큰 차이가 없었다. 덧붙여 가지치기 파라미터는 디폴트로 그 강도는 75, 자식마디 최소 레코드 수는 2개로 설정하였다. 또한 가지치기에 관한 변수를 조절하기 위해 2번에 걸쳐 실험을 실행하였다.

그 결과 〈표 5〉에 나타난 바와 같이 가지치기를 75로 지정하고 실험한 결과 학습 데이터에서는 87.50%의 정확률을, 실험 데이터에서는 83.125%의 정확률을 보여 주었다. 75의 가지치기 파라미터를 20으로 변경하자 학습 데이터에 있어 정확률이 95.16%로 상당히 좋아졌고 에러률은 앞의 실험보다 훨씬 줄어들었다. 그러나 학습 데이터에서 분류 정확률을 실험하는 것이기 때문에 신뢰성이 떨어질 수 있다. 따라서 정확한 실험을 위해 학습 데이터

〈표 5〉 가지치기를 75로 지정하고 측정한 실험 결과와 20으로 지정한 실험결과

	train data		test data	
Correct	560	87.50%	133	83.125%
Wrong	80	12.50%	27	16.875%
Total	640	100%	160	100%

	train data		test data	
Correct	609	95.16%	130	81.25%
Wrong	31	4.84%	30	18.75%
Total	640	100%	160	100%

에서 생성된 규칙을 실험 데이터에 연결하여 실행하였다. 이 결과 정확률은 81.25%가 나타났으며 이것은 앞의 실험결과와 큰 차이가 없었다. 따라서 본 논문에서는 가지치기 파라미터를 75회 지정하였고, 자식마디 최소 레코드 수는 2개로 설정하였다. 그 결과 〈표 6〉에서와 같이 학습 데이터에서 생성된 C5.0 모델은 총 39개의 rule을 생성하였다. 실험 데이터 분류 결과, 총 160건 중 128건인 80.0%가 정확히 분류하였으며, 20.0%인 32건은 성공적으로 이루어지지 못했다.

의사결정나무를 이용한 실험에서도 신경망 기법과 마찬가지로 범주 6, [Psychiatry and Psychology] 부분에 있어 90%라는 수치를 보여주었다. 이것은 자질별로 나무의 분기점을 만들며 학습하는 의사결정나무의 특성 때문이라고 생각된다. 따라서 자질의 특성이 뛰어난 범주 6의 경우 18건이 분류되는 높은 결과가 나타나는 반면, 문서가 불균형적으로 공통자질을 많이 포함하고 있는 범주 8의 [Natural Sciences]는 20건 중 14건만 바르게 분류되었다. 이는 범주 8이 고유한 대표 자질도 포함하고 있지만, 다른 자질도 포함하고 있는 문서가 많기 때문이며 불균형적인 분포를 보이고 있기 때문인 것으로 판단된다. 이것은 C5.0 알고리즘이 다른 범주의 자질을 자유로이 포함하고 있는 문서는 분류작업을 잘 수행하지 못하는 것을 보여준다.

반면 범주 6의 [Psychiatry and Psychology]의 분류 효율성이 높은 이유는 모든 문서에서 자질이 골고루 선정되었을 것이라고 추측할 수 있다. 그 예로 범주 6에는 schizophrenia, nervosa, anorexia, antipsy chotic 등과 같은 특성화된 자질들이 골고루 문서에 포함되어 있었다. 이 대표 자질들은 다른 범주에는 포함되지 않는 것으로 의사결정나무 규칙

〈표 6〉 C5.0의 분류 결과

범주명	test 문서 수	적중한 문서 수	Precision(%)
Anatomy	20	16	80
Organisms	20	17	85
Diseases	20	18	90
Chemicals & Drugs	20	14	70
Analytical, Diagnostic and Therapeutic Techniques and Equipment	20	15	75
Psychiatry and Psychology	20	18	90
Biological Sciences	20	16	80
Natural Sciences	20	14	70
총 테스트 문서 수	**160**	**128**	

Correct	128	80.00%
Wrong	32	20.00%
Total	160	

- 대상 : test data(160건)
- 정확도 : 80.00%
- 입력 레이어 : 360개의 뉴런
- 출력 레이어 : 8개의 범주

을 생성할 때 효율적인 분류작업을 가능하도록 하였다. 물론 범주 6도 다른 범주와 공통적으로 존재하는 자질을 가지고 있었다. 그 예로 Alzheim, stressl, asthenia 등의 자질은 [Diseases]의 자질과 공통적으로 사용되었다. 하지만 공통적으로 사용되는 자질 역시 편중되어 나타나는 것이 아닌 골고루 분포되어 있었고 기계학습 알고리즘에 큰 영향을 줄 수치는 아니라고 판단된다. 또한 분류를 실험할 실험 데이터의 문서가 범주 6만의 특정 자질들을 골고루 포함하고 있었기 때문에 다른 범주에 비해 분류효과가 탁월했을 것이라고 추측된다.

4.2.2 CHAID를 이용한 문서분류

CHAID는 의사결정나무에 기초한 분류 및 예측 모델이다. 가지 분리 방식은 가지 분리를 지원하며, 과적합의 문제없이 복잡한 모델을 생성할 수 있기 때문에 정확한 분류가 가능하다. 분류 실험은 디폴트 방식으로 수행하였다.

CHAID의 분류 결과는 〈표 7〉에 나타난 바와 같이 160건의 문서 중 76건이 분류된 47.70%의 정확률을 나타내었고, 84건인 52.50%는 성공적으로 이루어지지 못했다. 분류 결과 다른 알고리즘에서도 높은 분류 효율을 보였던 범주 6의 [Psychiatry and psychology]가 60%라는 분류효율을 나타내었다. 하지만 CHAID의 분류 결과가 다른 실험 가운데 가장 좋지 않은 실험 결과였다. 이번 실험 결과는 같은 데이터에 대해 어떤 기계학습을 적용하는가에 따라 학습의 결과가 많이 달라질 수 있음을 알게 해준 실험이었다.

4.3 신경망과 C5.0을 이용한 문서분류

신경망 기법을 이용한 분류 결과와 C5.0을 이용하여 나온 분류 결과를 연결하여 실험을

〈표 7〉 CHAID의 분류 결과

범주명	test 문서 수	적중한 문서 수	Precision(%)
Anatomy	20	9	45
Organisms	20	10	50
Diseases	20	11	55
Chemicals & Drugs	20	9	45
Analytical, Diagnostic and Therapeutic Techniques and Equipment	20	9	45
Psychiatry and Psychology	20	12	60
Biological Sciences	20	9	45
Natural Sciences	20	7	35
총 테스트 문서 수	**160**	**128**	

Correct	76	47.70%
Wrong	84	52.50%
Total	160	

- 대상 : test data(160건)
- 정확도 : 47.70%
- 입력 레이어 : 360개의 뉴런
- 출력 레이어 : 8개의 범주

수행하였다. 신경망 기법과 C5.0 기법의 분류 정확률 도출은 결합 모형이 아닌 병행모드의 결과라 할 수 있다. 즉 신경망 기법으로 학습 데이터를 훈련시켜 네트워크 가중치를 생성하고, C5.0을 사용하여 생성된 학습데이터의 정보 획득량을 산출하였다. 여기서 각 기법의 훈련결과를 실험 데이터에 병행 적용하여 분류결과를 도출하였다. 그 결과 〈표 8〉에 나타난 바와 같이 총 160건의 문서 중 134건인 83.75%가 정확히 분류되었으며, 26건인 16. 25%가 오분류되었다.

이는 신경망 기법과 C5.0 기법을 병행하여 사용된 분류가 단일 기법을 사용했을 경우보다 정확률이 약간 상승하였다. 즉, 신경망 기법은 전체 81.25%에서 2.5%가, C5.0의 경우 80.00%에서 3.75%가 상승되었다. 이것으로 보아 단일 문서분류보다 두 가지 분류기를 병행하여 사용하는 것이 근사한 차이지만 분류 효율이 향상되었다고 할 수 있겠다.

범주별 정확률을 살펴보았을 때, 앞의 신경망 기법과 C5.0 기법과의 두드러지는 차이점은 범주 1의 [Anatomy]가 가장 뛰어난 효율을 보인다는 것이다. 신경망 기법에서 뛰어난 효율을 보였던 범주는 [Analytical, Diagnostic and Therapeutic Techniques and Equipment]와 [Psychiatry and Psychology]이였으며, C5.0은 [Diseases]와 [Psychiatry and Psychology]이었다. 하지만 병행기법에서는 [Anatomy]의 분류효율이 100%로 가장 높게 나타났다. 일단 [Anatomy] 범주는 일정량의 특정 자질과 공통 자질을 포함하고 있는 범주로 실험 데이터 역시 자질의 고른 분포를 보여주고 있다. 하지만 이것은 특성화된 자질로 분류효율이 높아졌다고 판단할 수 없다. 왜냐하면 특정 자질 선정은 범주 6 [Psychiatry and Psychology]이 더 효율적

〈표 8〉 신경망 기법과 C5.0을 이용한 분류 결과

Correct	134	83.75%
Wrong	26	16.25%
Total	160	

- 대상 : test data(160건)
- 정확도 : 83.75%
- 입력 레이어 : 360개의 뉴런
- 출력 레이어 : 8개의 범주

범주명	test 문서 수	적중한 문서 수	Precision(%)
Anatomy	20	20	100
Organisms	20	17	85
Diseases	20	18	90
Chemicals & Drugs	20	17	85
Analytical, Diagnostic and Therapeutic Techniques and Equipment	20	15	75
Psychiatry and Psychology	20	19	95
Biological Sciences	20	14	70
Natural Sciences	20	14	70
총 테스트 문서 수	**160**	**134**	

으로 이루어졌기 때문이다.

그래서 그 원인을 파악하고자 C5.0의 rule set을 살펴본 결과 추측할 수 있었다. C5.0은 총 39개의 Rule을 생성했고 [Anatomy] 범주에 대해 7개의 rule을 생성했다. 이것은 범주 6이나 범주 7에 비해서 3-4개의 rule이 더 생성된 것이며, [Anatomy]범주의 적절한 자질의 선정과 함께 rule 적용 시 최대의 분류효율이 생성됐다고 판단된다. rule 생성에 관한 부분은 C5.0에서 여러 파라미터의 조절로 생겨나는 것이며, 최대의 분류효율을 나타내는 선에서 입력되었다. 따라서 [Anatomy] 범주의 7개의 rule이 분류의 정확률 향상에 도움을 주었다고 추측할 수 있다.

4.4 KNN을 이용한 문서분류

KNN 알고리즘은 다른 기계학습 기반 자동 분류 알고리즘에 비해 비교적 간단하다. 새로이 분류될 입력 문서가 있을 때 시스템은 학습 문서 집단 중에서 K개의 최 근접 문서를 찾아내며, K개의 최 근접 문서들이 할당된 범주 정보를 이용하여 후보 범주에 가중치를 부여한다. 본 논문에서는 K값을 3에서 7사이로 조종하여 실험을 수행하였으며 결과는 가장 수행결과가 좋았던 K값이 5일 때의 결과 값이다. 160건의 실험 데이터에 대한 실험결과는 〈표 9〉와 같다.

160건의 test data 중에 분류가 성공적으로 이루어진 data는 123건으로 76.875%에 해당된다. 그리고 남은 37건인 23.125%는 분류가 성공적으로 이루어지지 못했다. 가장 높은 정확률을 가지는 것은 [Psychiatry and Psychology]이다. 범주 6은 20건의 data 중에서 각각 18건을 분류하였다. 그 뒤로 범주 5 [Analytical, Diagnostic and Therapeutic

〈표 9〉 KNN의 분류 결과

범주명	test 문서 수	적중한 문서 수	Precision(%)
Anatomy	20	14	70
Organisms	20	14	70
Diseases	20	16	80
Chemicals & Drugs	20	16	80
Analytical, Diagnostic and Therapeutic Techniques and Equipment	20	17	85
Psychiatry and Psychology	20	18	90
Biological Sciences	20	15	75
Natural Sciences	20	13	65
총 테스트 문서 수	**160**	**123**	

Correct	123	76.875%
Wrong	37	23.125%
Total	160	

- 대상 : test data(160건)
- 정확도 : 76.875%
- 입력 레이어 : 360개의 뉴런
- 출력 레이어 : 8개의 범주

Techniques and Equipment]가 17건 분류로 정확률이 높았다.

앞의 두 실험과 마찬가지로 [Psychiatry and psychology] 범주가 다른 범주에 비해 정확률이 높은 것을 알 수 있다. KNN은 자질들의 Vector Distance 만으로 분류작업을 수행하기 때문에 자질의 특성이 중요한 변수가 된다고 추측할 수 있다. 범주 6 [Psychiatry and psychology]은 90%라는 정확률을 보여주었는데 이것은 일단 자질의 특성이 뛰어날 뿐만 아니라 다른 범주에서 범주 6의 자질을 포함하고 있기 때문인 것으로 판단된다. 마지막 이 실험은 분류 알고리즘에 있어 자질집합의 효율적인 선정이 정확률과 직결된다는 점을 알 수 있으며, 특성화된 자질의 선정은 다른 종류의 분류기로 실험을 해도 그 정확률이 항상 높다는 것을 알 수 있었다.

4.5 실험결과 분석

각기 다른 4가지 알고리즘을 가지고 실험한 결과를 살펴보면 총 160건에 문서에 대한 정확률은 신경망 기법과 C5.0 기법을 병행하여 사용하였을 때 83.75%로 가장 높았다. 신경망 기법만 사용했을 경우 81.25%의 정확률을 나타내었고, C5.0만 사용하였을 경우 80.00%를 나타냈다. 이것은 신경망 기법을 사용했을 때 나타난 81.25% 보다 2.5% 상승된 수치이며, C5.0만을 사용한 80.00% 보다 3.75% 상승된 수치이다. 따라서 두 기법을 같이 적용했을 경우 단일 알고리즘을 사용했을 때보다 분류 효율이 다소 높다고 할 수 있다. CHAID는 C5.0과 같은 의사결정분석 기법 중 하나지만 160건의 문서 중 76건의 분류 적중으로 47.70%의 효율을 나타냈다. 이것은 같은 의사결정 분석 기법이라도 어떤 알고리즘을 적용하느냐에 따라 효율성이 달라질 수 있다는 것을 의미한다. KNN은 160건의 문서 중 123건을 분류함으로 76.875%의 분류효율을 나타내었다(〈표 10〉 참조).

분류 결과의 정확률을 범주별 비교해 보면 가장 높은 정확률을 보이는 범주는 5개의 실험 결과 평균적으로 범주 6인 [Psychiatry and psychology] 것으로 나타났다. 이것은

〈표 10〉 분류 결과의 정확률 비교

실험 방법	실험 문서	적중한 문서 수	정확률	분류 효율이 제일 높은 범주	분류 효율이 두 번째로 높은 범주
Neural Network	160건	130건	81.25%	[5], [6] : 90%	[4] : 85%
C5.0		128건	80.00%	[3], [6] : 90%	[2] : 85%
CHAID		76건	47.70%	[6] : 60%	[3] : 55%
Nnet + C5.0		134건	83.75%	[1] : 100%	[6] : 95%
KNN		123건	76.875%	[6] : 90%	[5] : 85%

자질집합의 자질과 분류하려는 문서의 일치도가 높았기 때문이라고 분석된다. 또한 범주 6에는 특성화된 자질의 학습이 충분히 이루어졌으며 학습 데이터와 실험 데이터 간의 가중치 차이 역시 크지 않았기 때문에 분류 작업 시 그 효율이 높았다고 판단된다. 즉 다른 범주에는 포함되지 않는 특성화된 색인어가 골고루 포함된 것이 다른 범주에 비해 탁월한 정확률을 나타냈다고 추측할 수 있다. 또한 특정 분류 목록에서 정확률이 떨어지는 현상을 분석해 본 결과, 자질집합의 자질과 분류하려는 문서에 존재하는 단어의 일치도가 매우 낮음을 발견하였다. 또한 [Organisms]와 [Natural sciences]가 낮은 분류 성능을 보인다는 것을 알 수 있다.

[Natural Sciences]가 낮은 분류 효율을 보이는 이유는 몇 가지로 추측할 수 있다. 먼저 자질의 문제이다. 여기에는 이 범주에 속하는 문서이긴 하지만 그 자질은 다른 범주에 공통적으로 속하는 경우가 많아 특성화된 색인어가 추출되지 못했기 때문일 수 있다. 또 다른 이유로 학습 데이터와 실험 데이터 사이에 자질 집합의 가중치 차이에 기인할 수 있다. 즉 학습 데이터에서는 중요하다고 판단되지 않던 자질이 실험 데이터에 포함되어 있었고, 이 부분에 대한 학습이 충분히 이루어지지 않았기 때문에 분류 작업은 효율적으로 이루어지지 않았다고 판단된다.

범주별 정확률을 살펴보았을 때, 앞의 신경망 기법과 C5.0 기법을 동시에 사용하였을 때 [Anatomy]가 가장 뛰어난 효율을 보였다. 그 이유로는 C5.0는 39개의 Rule 중 [Anatomy] 범주에 대해 7개의 Rule을 생성하였다는 것이다. 이는 다른 범주들에 비해 3-4개가 많은 Rule을 생성한 것이며 이로 인해 자질의 고른 분포와 많은 Rule의 생성으로 100%라는 분류 효율이 나타났다고 판단할 수 있다.

종합해 본 결과, 낮은 분류 효율을 가지는 범주가 나타나는 이유는 자질 선정의 문제라고 판단된다. 훈련 자질과 실험 자질간의 가중치 차이가 있었기 때문에 학습 프로그램이 충분히 학습하고 규칙을 생성하지 못했다. 또한 이 범주만의 특성화된 자질이 생성되지 못했다. 이 범주를 대표할 수 있는 특성화된 자질이 생성되어 있었어도 실험 데이터에 포함되지 못해 분류의 효율성이 낮게 측정되었다.

즉, 자질 집합의 신뢰성 및 실험 데이터의 객관성 유지도 분류 성능에 많은 영향을 끼친다는 것을 알 수 있었다. 마찬가지로 보다 좋은 결과를 가져오기 위해서는 중요도가 높고 신뢰성 있는 자질의 추출이 중요하다고 할 수 있다. 덧붙여 광범위한 주제 분야의 경우 좀 더 정확하고 많은 문서의 수집이 이루어져 분류의 성능을 높여야 할 것으로 보인다.

5. 결론

문서의 자동분류는 사실상 정확한 분류가 어렵다. 하지만 사람의 손을 거치지 않고 자질

을 추출한 후 빠른 방법으로 분류를 수행하여 정확률을 나타내는 것은 웹상에서 응용 가능하다는 점에서 유용하다.

본 논문에서는 수작업 분류 시 초래하는 여러 가지 한계점을 극복하고, 이용자에게 보다 빠르고 정확한 분류 서비스를 제공하기 위해 기계학습 알고리즘을 도입하였다. 연구대상으로는 MeSH의 8개의 주제별 범주로 각각 100개의 문헌 타이틀을 선정하였으며, 4개의 기계학습 알고리즘으로 5개의 분류 실험을 수행하였다. 본 논문은 이용자들이 필요로 하는 의학 자료를 방대한 양의 데이터 안에서 선택적으로 추출하게 함으로 보다 효율적인 분류 서비스를 수행하게 하는데 그 목적이 있다.

신경망 기법은 추론능력과 분류 능력에는 탁월하며 비교적 특성이 뚜렷하지 않는 데이터에 대해서도 학습을 수행하는 특성이 있지만, 분류 과정에서 일어나는 규칙에 대해서 명확히 설명할 수 없는 한계를 가지고 있다. 이에 비해 C5.0은 분류과정의 명확한 설명이 가능하나 가지치기 등의 문제로 과 적합의 문제가 생길 수 있다. 따라서 신경망 기법의 네트워크 가중치와 C5.0의 Rule set을 사용한다면 분류과정에 관한 설명 및 과적합의 문제가 정도는 감소할 수 있다. 마찬가지로 C5.0으로 귀납학습 후 신경망 기법을 사용한다면 보다 효율적인 정확률을 나타낼 수 있다.

본 논문에서는 신경망 기법과 C5.0 기법을 병행하여 사용했을 경우 단일 기법을 사용했을 경우보다 2.5%, 3.75%가 상승하여 분류 효율이 83.75%로 측정되었다. 이 수치는 5개의 분류 실험 결과 중 가장 높은 정확률을 나타내었으며, 기존의 연구와 비교했을 때 크게 변화가 없다. 기존 연구(김천식 2006; 김성희 2000)에서 알 수 있듯이 신경망 기법과 C5.0 기법의 병행 사용은 신경망 기법이 분류과정을 명확히 설명하지 못하는 한계점과 C5.0의 과적합의 문제를 감소해주어 분류 오차를 줄여주는 역할을 수행한다고 추측할 수 있다.

앞으로 문서분류는 MeSH의 의학 자료뿐만 아니라 여러 주제 분야에 걸쳐 연구가 이루어져야 하며, 이용자들이 원하는 실제적인 주제에 대해 빠르게 구분하여 필요로 하는 정보를 찾을 수 있도록 해야 한다. 신경망 기법과 C5.0 기법의 병행 사용 시 단일 기법을 사용한 것 보다 정도의 문제점 해결로 정확률이 상승되지만 상승된 정확률은 근사한 차이이다. 따라서 두 모델간의 장점만을 취하여 분류할 수 있는 결합 알고리즘의 개발이 필요하다고 할 수 있다.

참고문헌

김성희. 2000. WWW상의 지능형 정보검색을 위한 기계학습 알고리즘 구현에 관한 연구. 『정보관리학회지』, 17(2): 189-205.

김진화. 2004. 인공지능을 이용한 웹 문서의 자동분류. 『서강경영논총』, 15(2):49-75.

김천식, 홍유식. 2006. 텍스트마이닝을 이용한 XML 문서분류 기술. 『한국 컴퓨터 정보학회 논문지』, 11(2): 19.

김판준, 이재윤. 2007. 문헌 간 유사도를 이용한 자동분류에서 미 분류 문헌의 활용에 관한 연구.『한국정보관리학회지』, 3: 251- 271.

노영희. 2001. 기계학습을 기반으로 한 인터넷 학술문서의 효과적 자동분류에 관한 연구.『한국도서관정보학회지』, 32(3): 307-330.

박성배, 태윤식. 2007. 기계학습과 정보검색.『정보과학회지』, 25(3): 5-11.

이영숙, 정영미. 2001. 계층적 분류체계를 위한 자동분류 기법에 관한 연구.『한국정보관리학회 학술대회 논문집』, 173-176.

이재문. 2003. 휴리스틱을 이용한 KNN의 효율성 개선.『정보처리학회논문지』, 10-B(6): 719-720.

조용준 외 3인. 1999.『Neural Connection을 이용한 데이터마이닝 신경망 분석』. 서울: SPSS 아카데미.

최종후 외 3인. 2000.『AnswerTree를 이용한 데이터마이닝 의사결정나무분석』. 서울: SPSS 아카데미.

허명회, 이용구. 2003.『데이터마이닝 모델링과 사례』. 서울: SPSS 아카데미, 43.

허인옥. 2002.『신경망 학습을 이용한 문서 자동분류』. 석사학위논문, 성신여자대학교 교육대학원.

황성하, 최광남, 이대규, 이상호. 2005. 인터넷 문서의 자동분류 서비스 시스템에 관한 구현.『한국 컨텐츠학회 추계종합학술대회 논문집』, 3(2): 66-71.

SPSS Korea 컨설팅 팀. 2007.『Clementine Manual』. 서울: SPSS 코리아.

Chang, Jeffrey, 2000. "Using the MeSH Hierarchy to Index Bioinformatics Articles." CS224N/Ling237 Final Projects 2000, 1-10.

Cherkassky. V and N. Vassilas. 1989. "Performance of Back Propagation Networks for Associative Database Retrieval." Proc. International Joint Conference on Neural Networks, 1: 77-84.

Chidanand Apte, Fred Damerau, and Sholom M. Weis. 1994. "Towards language independent automated learning of text categorization models." In Proceeding of the 17th annual international ACM-SIGIR, 23-30.

Jacobs. P. 1993. "Using statistical methods to improve knowledge-based news categorization." IEEE Expert, 1-10.

Lewis, D. D. 1992. Representation and Learning in information Retrieval Ph.D. thesis. Department of Computer Science. University of Massachusetts. Amherst.

Ruiz, Miguel E. and Padmini Srinivasan. 2002. "Hierarchical Text categorization Using Neural Networks." Information Retrieval, 5(10):87-118.

Salton G. and M. J. McGill. 1983. An Introduction to Modern Information Retrieval. New York: McGraw-Hill.

Salton, G. E. A. Fox and H. Wu. 1983. "Extend boolean information retrieval." Communications of the ACM, 26(12): 1022-1036.

Tom Mitchell. 1996. Machine Learning. New York: McCraw Hill.

Ruiz, M. E. and Padmini Srinivasan. 2002. "Hierarchical Text categorization Using Neural Networks." Information Retrieval, 5(10):87-118

Vapnik, V. 1995. The Nature of Statistical Learning Theory. New York: Springer.

Yang, Y. and Xin Liu, 1999. "A re-examination of text categorization methods". Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 42-49.

Yang, Y. 1999. "An evaluation of statistical approaches to text categorization." Journal of Information Retrieval, 1(1-2): 69-90.

Yang, Y. J. O. Pederson. 1997. "A Comparative study on feature selection in text categorization." In Proceeding of the 24th International Conference on Machine Learning, 412-420.

Yang, Y. L. Xin. 1999. "A re-examination of text categorization methods." Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, 42-49.

Yang, Y. 1994. "Expert Network: Effective and efficient learning from human decisions in text categorization and retrieval", In 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 13-22.

Yang, Y. Pederson, J.O. 1997. "Feature selection in statistical learning of text categorization."Proceedings of the 14th International Conference on Machine Learning, 412-420.