퍼지추론과 신경망을 사용한 유즈넷 뉴스그룹 결정

Determination of Usenet News Groups by Fuzzy Inference and Neural Network

김종완*, 김희재*, 김병만**

* 대구대학교 컴퓨터·IT공학부

** 금오공과대학교 컴퓨터공학부

Jong-Wan Kim*, Hee-Jae Kim*, and Byeong Man Kim**

* School of Computer and Information Technology, Daegu University

** School of Computer Engineering, Kumoh National Institute of Technology

E-mail : jwkim@daegu.ac.kr

요 약

본 연구에서는 다양한 뉴스그룹들 중에서 사용자의 취향과 유사한 뉴스그룹들을 코호넨 신경망을 이용하여 추천해주는 방법을 제시한다. 신경망을 학습시키기 위한 뉴스 문서의 키워드들을 선택하기 위해 예제 문서들로부터 후보 용어들을 추출하고 퍼지 추론을 적용하여 대표 용어들을 선택한다. 하지만 신경망의 학습패턴을 관찰해 보면, 많은 부분이 비어있는 희소성 문제를 발견할 수 있다. 이에 본 연구에서는 통계적인 결정계수를 도입하여 불필요한 차원을 제거한 후 신경망을 학습시키는 새로운 방법을 제안한다. 제안된 방법은 모든 차원을 활용할 때 보다 클러스터내 거리와 클러스터간 거리의 척도를 이용한 클러스터 중첩도 면에서 우수한 분류 성능을 보여줌을 확인하였다.

1. 서론

사용자 프로파일과 부합되는 뉴스그룹을 결정하기 위해서는 각 뉴스그룹을 대표하는 용어들을 추출하고 이들을 사용자 프로파일과 비교하여 유사성이 높은 뉴스그룹들을 선택하는 일들이 필요하다. 뉴스그룹을 대표하는 용어들을 추출하기 위해, 인터넷에 접속해서 수집한 뉴스들로부터 후보 용어들을 추출하고 퍼지추론을 적용하여 대표 용어들을 선택하였다. 제안 방법의 성능은 대표 용어들을 선택하는 방법에 의해 영향을 크게 받으며, 또한 뉴스그룹에서 대표 용어를 추출하는 문제 자체가 불확실성을 내포하고 있어, 이러한 문제 해결에 효과적인 퍼지추론을 이용한 대표 용어의 선정 방법을 채택하였다.

사용자 프로파일과 각 그룹의 대표 용어들과의 유사성을 판단하기 위한 방법으로는 정보검색 분야에서 널리 쓰이는 코사인 유사도 방법[1], 신경망을 이용한 방법 및 기타 학습 기법을 이용한 방법들을 쓸 수 있다. 본 논문에서는 다루기 쉽고 성능도 우수한 신경망을 이용한 방법을 선택하였다.

코호넨 신경망(Kohonen network)은 교사의 지시 없이 뉴스그룹 문서들로부터 추출된 키워드 즉 대표 용어만 가지고 자연스럽게 뉴스그룹간의 연관 관계를 찾을 수 있다는 장점이 있다. 이에 본 연구에서는 코호넨 신경망을 학습 알고리즘으로 채택하였다. 하지만, 신경망의 훈련벡터로 사용되는 패턴을 관찰해 보면, 많은 뉴스그룹에서 선택된 특정한 키워드부분이 비어있는 희소성 문제를 발견할 수 있다. 이러한 희소성 문제를 해결하기 위해, 본 논문에서는 사용자가 제시하는 목표변수(즉 유사한 뉴스그룹)와 관련성이 높은 입력변수(여기서는 선택된 대표 용어)를 선정하

여, 이를 기준으로 학습시키는 것이 입력변수의 전체 차원을 함께 학습시키는 것보다 유용할 것이라는 판단 하에 통계적인 방법을 도입하였다.

2. 제안된 뉴스그룹 자동 결정 방법

2.1 대표 용어 선택 방법

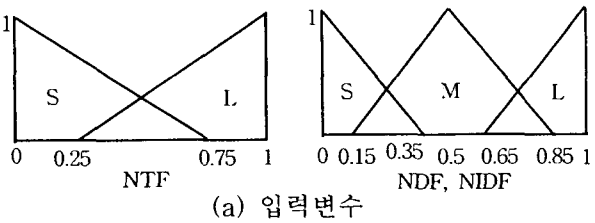
뉴스그룹의 예제 문서들로부터 뉴스그룹을 가장 잘 대변하는 대표 용어의 선택은 중요하다. 문서 집합에서 대표 용어를 추출하고 이들의 가중치를 부여하는 문제는 기존의 대표적인 선형 분류기인 Rocchio와 Widrow-Hoff 알고리즘들[2]이 학습 문서 집합을 대표하는 중심 벡터를 구성하는 것과 성격이 동일하다. 이들 알고리즘들은 용어의 가중치 산정시 발생 빈도수(TF)와 역문헌 빈도수(IDF)를 결합하는 방법을 취하고 있지만, 문서내 또는 문서 집합내 용어들간의 관련성을 용어의 가중치 계산에 반영하고 있지는 않다.

따라서 TF가 높은 용어는 높은 가중치를 가지게 되는데 대표 용어로서 실제 중요하지 않는 용어임에도 문서 내에 자주 발생만 되면 높은 가중치 값을 부여받을 수 있다는 단점을 지니고 있다.

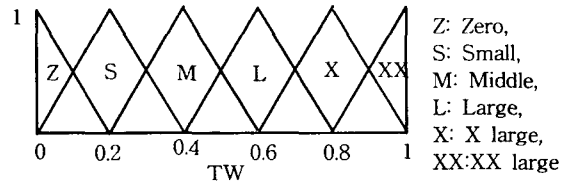
이러한 문제를 해결하기 위해, 특정 용어의 중요도 계산에 사용되는 입력 정보(예: TF, IDF)들은 정량적으로 정확히 해석될 수 없는 부정확하고 불확실한 특성을 내포하고 있으므로, 이러한 불확실성의 문제 해결에 효과적인 퍼지추론을 적용하여 후보 용어들의 가중치를 계산하고 이 값들에 따라 선택 우선순위를 부여하는 방법도 있다[3].

퍼지추론을 이용한 대표 용어 중요도를 계산하기 위해 뉴스 문서들은 불용어 처리 과정을 거치고, Porter stemmer를 사용하여 영문 단어를 추출하고 대표적인 한글 형태소 분석기[4]를 사용하여 한글 명사를 추출하여 이를 후보 용어들의 집합으로 변형하며, 이 집합으로부터 각각의 용어들의 TF(Term Frequency), DF (Document Frequency), IDF(Inverse Document Frequency) 정보가 구해진다. 이들 정보들이 퍼지추론을 위한 퍼지시스템의 입력으로 이용된다.

그림 1은 퍼지 추론을 위하여 사용된 입출력 변수들의 멤버쉽 함수를 나타내고 있다. 또한 표 1은 NTF 퍼지 입력값의 소속 정도에 따라 두 부분으로 나누어 규칙들을 표현하고 있다. NTF, NDF, NIDF 퍼지 입력값을 위의 결과로 생성된 18개의 추론 규칙별로 이들의 전건부의 소속 함수에 적용시킨다. 각각의 소속 정도가 구해지면 이들 중에서 최소값을 취한다. 그 결과 규칙별로 하나씩의 퍼지 값이 생성되며 이 퍼지 값들을 퍼지 출력변수 TW에 따라 6개의 그룹으로 분류하고 그룹별로 해당 그룹에 속한 퍼지 값들 중 최대값을 취하여 총 6개의 퍼지 값들을 생성한다. 최종적으로 이들 6개의 퍼지 값들을 무게중심법으로 비퍼지화한 값이 해당 용어의 중요도 값으로 결정된다.



(a) 입력변수



(b) 출력변수

그림 1 퍼지 입출력변수

NDF \ NIDF	S	M	L
S	Z	S	M
M	S	L	X
L	S	X	XX

NTF = S

NDF \ NIDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NTF = L

표 1 퍼지 추론규칙

2.2 결정계수를 이용한 학습패턴의 차원 축소

학습 작업에서 어느 속성이 클래스를 예측하는데 기여하는지 결정하는 방법들로 속성들 모두에 걸쳐서 평균 유사성 척도를 계산하는 근사 이웃(nearest neighbor) 알고리즘들이 제안되었다. 하지만 단순한 근사 이웃 알고리즘들은 모든 속성들을 동일한 가중치로 판단하므로 속성들 사이의 패턴 분류 기여도를 적절하게 산정하지 못하였다. 이를 해결하기 위해 여러 가지 가중치를 부여하는 방식도 제안되었다[5]. 이 방법은 일종의 주성분 분석(PCA:Principal Component Analysis) 기법으로서, 낮은 특성값을 갖는 차원을 삭제하는 Singular Value Decomposition(SVD) 방법을 채택하고 있다.

주성분 분석 기법은 원 변수들(y_i)의 선형 결

합으로 이루어지는 주성분 예측변수들($\hat{y}_i$)를 구해서, 이 변환된 변수들을 패턴분류에 사용하는 것이므로 어떤 입력 성분이 패턴 분류에 기여하는 지 알 수는 없다. 하지만 본 연구에서는 특정 성분이 패턴 분류 학습에 필요한 것인지 결정하는 것이 중요하므로, 이러한 방법보다는 패턴들의 분류 기여도를 결정해주는데 유용한 통계학의 결정계수(coefficient of determination) R^2를 사용하였다[6].

먼저 결정계수를 이용해서 뉴스그룹 데이터를 분류하려면 목표변수가 있어야 한다. 따라서 본 연구에서는 목표변수인 뉴스그룹의 클래스를 뉴스그룹 도메인 이름 기준으로 지정하였다. 예를 들면, NNTP 서버인 news.kornet.net 에 있는 126개의 뉴스그룹을 영역기준으로 분류하였다. 즉 han.answers.all을 클래스 1, han.arts.architecture.all을 클래스 2, 나머지도 이런 식으로 분류하니 모두 114개의 그룹이 나왔다.

이 114개의 클래스 변수를 목표변수로, 목표변수들과 관련성이 있는 후보용어들을 입력변수로 두고, 모든 후보용어들에 대하여 목표 클래스 변수와의 결정계수를 계산한다. 계산된 결정계수의 값 중에서 미리 지정한 임계치 이하인 가장 낮은 결정계수의 값을 갖는 후보용어를 하나씩 제거하는 후진소거(backward elimination) 변수선택법[6]을 수행하여 불필요한 용어를 제거하였다.

3. 실험 및 분석

3.1 실험 데이터 수집 및 학습 방법

본 논문에서 제안된 방법은 자바 언어로 구현되었다. 먼저, 훈련 데이터를 수집하려고 자바의 java.net.Socket 클래스를 이용하여 유즈넷 뉴스서버인 news.kornet.net에 접속한 후, NNTP 프로토콜을 통해서 뉴스그룹을 선택하고 각 뉴스그룹에서 뉴스 문서를 내려 받았다.

실험은 126개의 뉴스그룹을 대상으로 하였으며, 퍼지추론으로 대표 용어를 추출하는 경우에 뉴스그룹당 20개의 문서를 임의로 추출한 경우를 실험하였다. 출력뉴런의 크기는 5*5로 정하였으며, 훈련은 1,000회 실시하였다. 훈련 데이터는 각 뉴스그룹에서 퍼지추론으로 추출하고 결정계수를 적용해서 일부 연관도가 낮은 성분을 제거한 용어들을 데이터베이스에 저장해 놓고, 각 뉴스그룹의 문서에서 용어들을 분석한다. 본 논문에서는 126개 뉴스 그룹에서 28개의 단어를 추출하여 사용하였다.

각 뉴스그룹의 문서의 수와는 상관없이 단어의 개수만을 파악했을 경우 문서의 수가 많은 뉴스그룹에서는 대체적으로 단어의 빈도수가 많다. 예를 들어, "han.comp.os.linux.networking" 뉴스그룹의 경우 문서의 수가 1448개인 반면, "han.answers" 뉴스그룹은 24개의 문서만 데이터베이스에 저장되어 있다. 이런 편차를 줄이기 위하여 본 논문에서는 정규화(normalization)를 수행하였다.

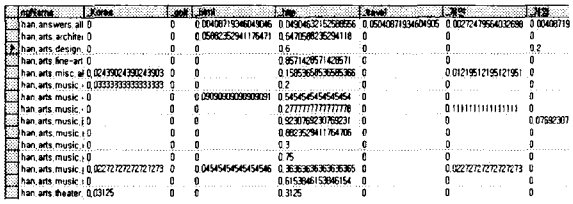


그림 6 126개 뉴스그룹의 정규화된 입력벡터

3.2 결정계수 도입 효과분석

차원을 축소하기 위한 결정계수의 효과를 살펴보기 위하여, 차원을 감소시켰을 때와 그렇지 않을 때의 클러스터내 거리(Dw) 및 클러스터간 거리(Db)를 아래와 같이 정의하고 계산하였다.

$$Dw_j = \frac{1}{|C_j|} \sum_{i \in C_j} \sqrt{[X_i - W_j]^2} \qquad (1)$$

여기서 X_i는 클러스터 j에 속하는 i번째 학습패턴을, W_j는 j번째 출력뉴런의 연결강도벡터 즉, j번째 클러스터의 중심벡터를 의미하고, C_j는 j번째 클러스터에 속하는 패턴들의 집합을, $|C_j|$는 j번째 클러스터에 속하는 패턴들의 수를 나타낸다. 따라서 이들 간의 거리인 Dw_j는 j번째 클러스터의 중심벡터와 학습패턴간의 거리를 의미하며, 이를 모든 출력뉴런들의 합으로 정의하여 전체 클러스터의 수로 나눈 아래의 식은 클러스터내 거리(Dw)를 나타낸다.

$$Dw = \frac{1}{k} \sum_{j=1}^{k} Dw_j \qquad (2)$$

여기서 k는 출력뉴런들의 수, 즉 클러스터의 개수를 나타낸다.

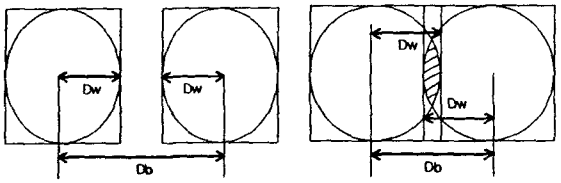
클러스터간 거리 Db는 식 (3)과 같이 각 클러스터별로 j번째 클러스터 자신을 제외한 다른 클러스터들과의 거리(Db_j)를 계산하고 이를 평균한 식 (4)로 정의된다.

$$Db_j = \sum_{m=1, m \neq j}^{k} \sqrt{[W_j - W_m]^2} \qquad (3)$$

여기서 W_j와 W_m은 각각 j번째 뉴런과 m번째 뉴런의 연결가중치를 나타내므로, 이들 간의 거리는 클러스터들 사이의 거리를 의미한다.

$$Db = \frac{1}{k} \sum_{j=1}^{k} Db_j \qquad (4)$$

일반적으로 좋은 패턴 분류기는 클러스터내 거리는 줄이면서 클러스터간 거리는 늘리는 것이므로 이를 제안된 차원감축 방법의 성능을 평가하는데 활용한다[7].



case 1: 클러스터 분리 (2Dw ≤ Db)　　case 2: 클러스터 중첩 (2Dw > Db)

그림 3 클러스터 중첩도

표 2의 20개 문서를 대상으로 실험한 경우에는, 용어의 수가 28개이며, 0.01과 0.02의 결정계수 임계치를 사용할 때, 각각 32%와 43%의 용어 수를 줄일 수 있었다. 특히, 두 클러스터내 거리합인 2*Dw 보다 클러스터간 거리 Db가 작은 그림 3의 case 2에 해당하므로 클러스터들 간에 중첩이 있음을 알 수 있다. 그래서 클러스터 중첩도 계산이 어려운 원 대신에 정사각형으로 간주하고, 최대한 원의 면적과 유사하도록 중첩 사각형 면적의 절반을 계산하는 식 (5)에 따라 클러스터간의 중첩도를 계산한 결과, 제안된 방법들이 임계치에 상관없이 클러스터간 중첩도가 50% 이상 개선됨을 알 수 있었다.

$$\text{중첩도} = (2Dw - Db) * Dw \qquad (5)$$

사용된 용어수	기준방법 (28 용어)	임계치 0.01 (19 용어)	향상률 (%)	임계치 0.02 (16 용어)	향상률 (%)
Dw	0.4	0.35	12.5	0.36	10.0
Db	0.61	0.62	1.6	0.63	3.3
중첩도	0.076	0.028	63.2	0.0324	57.4

표 2 126 뉴스그룹의 20 문서 대상 실험결과

위의 실험 결과를 통해서 우리는 용어 수를 지나치게 많이 줄이는 것이 반드시 좋지는 않다는 사실과 제거되는 용어를 선정하는 방법이 중요하다는 것을 확인할 수 있었다. 이러한 결과는 기본적으로 제안된 방법이 입력 차원이 보다 많은 문제에 효과적이라는 사실을 입증하여 준다.

4. 결론 및 향후 과제

본 논문에서는 사용자 프로파일에 기반한 뉴스 리더의 주요 부분인 프로파일-뉴스그룹 맵핑 방법을 제안하고 이의 성능을 분석하여 보았다. 뉴스그룹의 문서를 대상으로 퍼지추론을 수행하여 뉴스문서를 대표하는 용어를 추출하였고, 결정계수를 도입하여 패턴 분류 기여도가 낮은 차원을 감축시켰으며, 선정된 용어를 클러스터링하기 적합한 코호넨 신경망으로 학습시켰다.

본 연구에서는 첫째, 퍼지추론을 통한 뉴스문서로부터 대표 용어들을 추출하여 보다 정확도를 높였다. 둘째, 학습에 불필요한 중복된 속성들을 제거하기 위하여 통계학의 결정계수를 활용하여 패턴 분류율을 향상시켰다. 셋째, 제안된 방법을 패턴 분류율 면에서 성능을 평가하기 위하여, 클러스터내 거리 및 클러스터간 거리의 척도 면에서 비교하였다. 특히 클러스터내 거리합이 클러스터간 거리보다 커지는 클러스터 중첩의 정도를 정의하고, 이를 기준으로 제안된 방법의 우수성을 확인하였다.

향후에는 입력벡터의 차원이 보다 큰 복잡한 문제에 적용시켜서 제안된 결정계수를 이용한 차원감소 효과의 유용성을 확장할 필요가 있다.

참고문헌

[1] G. Salton and M. McGill, Introduction to Modern Information Retrieval, New York, McGraw Hill, 1983.

[2] David D. Lewis, Robert E. Schapire and James P. Callan and Ron Papka, "Training algorithms for linear text classifiler", Proceedings of SIGIR-96, 1996.

[3] Byeong Man Kim, Ju Youn Kim and Jongwan Kim, "Query Term Expansion and Reweighting using Term Co-Occurrence Similarity and Fuzzy Inference," Proc. of IFSA/NAFIPS, pp.715-720, 2001.

[4] 한국어 형태소 분석기와 한국어 분석 모듈 (HAM: Hangul Analysis Module), http://nlp.kookmin.ac.kr/."

[5] Terry R. Payne and Peter Edwards, "Dimensionality Reduction through Sub-Space Mapping for Nearest Neighbor Algorithms," European Conference on Machine Learning, pp.331-343, 2000.

[6] 강현철, 한상태, 최종후, 김은석, 김미경, SAS Enterprise Miner 4.0을 이용한 데이터마이닝 - 방법론 및 활용, 자유아카데미, 2001.

[7] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.