신경망 또는 k-NN에 의한 신문 기사 분류와 그의 성능 비교

조태호
삼성 SDS, 정보 기술 연구소
서울시 강남구 역삼동 718-5 삼성 멀티 캠퍼스 13 층
전화. 02-3429-2732
FAX 02-3429-2800
Email. tcjo@unitel co kr

The Comparison of Neural Network and k-NN Algorithm for News Article Classification

Taeho Charels Jo
Information Technology R&D Center, Samsung SDS
13th F Jlok Bldg 707-19 Yoksamdong KangnamGu, Seoul
Tel· 02-3429-2732
FAX· 02-3429-2800
Email: tcjo@unitel co kr

요약

텍스트 마이닝(Text Mining)이란 텍스트형태의 문서들의 패턴 또는 관계를 추출하여 사용자가 원하는 새로운 정보를 가공하거나 기존의 정보를 변형하는 과정을 말한다 텍스트 마이닝의 기능에는 문서 범주화(Document Categorization) 문서 군집화(Document Clustering), 그리고 문서 요약(Document Summarization)이 이에 해당된다 문서 범주화란 문서에게 사전에 정의한 범주를 부여하는 과정을 말하고, 문서 군집화란 문서들을 계층적 구조로 형성하는 과정을 말하고 문서 요약이란 문서의 전체 내용을 대표할 수 있는 내용의 일부만을 추출하는 과정을 말한다 이 논문에서는 문서 범주화만을 다룰 것이며 그 대상으로는 신문 기사로 설정하였다 그의 범주는 4 가지로 정치, 경제, 스포츠, 그리고 정보통신으로 설정하였다 문서 범주화는 문서 분류(Document Classification)라고도 하며 문서에 범주를 자동으로 부여하여 기존에 인위적으로 부여함으로써 소요되는 시간과 비용을 절감하는 것이 목적이다 문서 범주화에 대하여 k-NN(k-Nearest Neighbor)와 신경망을 이용하였으며, 신경망을 이용한 경우가 k-NN 을 이용한 경우보다 성능이 우수하였다

1. 서론

컴퓨터 하드웨어 기술이 발전함에 따라 데이터를 저장하는 매체도 대용량화 되어가고 있다 그에 따라 저장된 데이터양도 기하급수적으로 증가하게 되었고, 사용자가 요구하는 정보도 다양해졌다. 또한 인공지능 기술의 발전으로 저장된 대용량의 데이터를 활용하여 이들의 패턴 또는 관계를 추출함으로써 새로운 정보를 가공하는 것이 대두되었다. 이러한 과정을 데이터 마이닝이라 한다 [1][2] 이를 접근하는 방법으로는 통계적 방법 이외에 신경망, 유전자 알고리즘, 퍼지 등이 있다 [3][4]

실제로 사용자는 정형의 수치 데이터 보다 비정형의 텍스트 데이터를 다루는 경우가 빈번하므로 수치 데이터를 다루는 데이터 마이닝인 경우 저장된 데이터의 활용도를 증가 시키는 데 한계가 따른다. 그리하여 수치 데이터뿐만 아니라 텍스트 데이터를 활용하여 사용자가 원하는 정보를 가공하는 것이 필요하게 되었다. 이의 과정을 텍스트 마이닝이라 한다 [6][7]

텍스트 마이닝이란 텍스트 형태의 문서의 패턴 또는 관계를 추출하여 새로운 정보를 가공하는 과정으로 문서 범주화, 문서 군집화, 그리고 문서 요약이 이에 해당된다 [6][7][8][9] 문서 범주화란 문서에 미리 설정된 범주를 부여하는 과정으로 문서 분류라고도 할 수 있다 [10] 문서 군집화란 내용이 유사한 문서들을 여러 개의 집단으로 구성하는 것을 말하며, 문서 요약이란 문서의 전체 내용을 대표하는 일부의 내용만을 추출하는 과정을 말한다

이 논문의 내용은 문서 범주화에 한정되고, 이의 목적은 문서를 범주별로 저장할 경우 이를 자동화 하여 이에 대한 비용을 절감하기 위함이다 이를 자동화 하기 위해 많은 연구가 진행되어 왔다 [14]에서는 Support Vector Machine 을 문서 범주화에 적용하는 것을 제안하고 있으며, [15]에서는 적은 수의 단어를 이용하여 문서를 계층적으로 분류하는 방법을 제안하였다. [16]에서는 비교사 모델(Unsupervised Model)과 교사 모델(Supervised Model)을 혼합하여 문서를 범주화 하는 방법을 제안하고 있고, [17]에서는 의학 분야의 문서에 대하여 k-Nearest Neighbor 와

베이지안 분류기를 혼합하여 이를 범주화 하는 방법을 제안하고 있다. [18]에서는 문서 범주화에 대하여 Widrow-Hoff 알고리즘과 Exponentiated Gradient 알고리즘의 두개의 학습 알고리즘을 제안하였다.

이 논문에서는 신문 기사 분류를 예로 하여 신경망과 k-NN(k-Nearest Neighbor)의 접근에 대한 성능을 비교하겠다 여기서는 신경망 모델로 역전파 알고리즘을 사용할 것이며, 이에 대한 자세한 설명은 [11] 또는 [12]를 참고하기를 바란다 k-NN 알고리즘은 어느 특정한 패턴에 대하여 학습 데이터로 저장된 패턴 중에서 거리가 최소인 k 개의 패턴을 선정하고 그 중 다수가 속한 범주를 주어진 패턴의 범주로 분류하는 알고리즘이다. 이에 대한 자세한 설명은 [13]을 참고하기를 바란다

이 논문의 구성에서 2 절은 문서의 비정형 데이터로부터 수치적으로 표현한 일정한 차원의 벡터로 추출하는 과정을 제시할 것이다. 제 3 절에서는 신문 기사 분류를 예로 하여 k-Nearest Neightbor 와 역전파 알고리즘의 성능 비교 결과를 제시하겠다. 마지막으로 4 절에서는 실험 결과의 전반적인 분석과 이 논문에서 제시한 방법의 개선점과 앞으로의 과제를 제시하겠다

2. 문서에 대한 특징 추출

이 절에서는 backdata 로 저장된 각각의 문서를 k-Nearest Neightbor 또는 역전파 알고리즘에 적용할 수 있는 수치 벡터로 전환하는 과정을 설명하겠다. 어느 개체로부터 분류에 결정적인 인자를 추출하는 과정을 특징 추출(Feature Extraction)이라 한다 [13]

문서의 특징 추출은 키워드를 속성으로 하고, 키워드에 대한 빈도수를 속성값으로 하여 벡터의 형태로 추출된다 [14][15][16][17][18]. 문서 범주에 대해서는 어느 특정한 범주의 문서에 대하여 집중적으로 등장하는 키워드가 문서의 특징으로서 추출 되어야 한다

우선 문서의 범주가 다음과 같이 주어져 있음을 가정하자.

$$C = \{c_1,\ c_2,\ \ldots..\ ,\ c_m\}$$

C 는 분서 범주의 집합이고, 그의 원소 c_i 는 문서의 범주를 나타낸다.

특징 추출을 위한 첫번째 과정으로 backdata 로 저장된 문서로부터 키워드와 각각에 대하여 전체 문서에 대한 총 빈도수와 범주별 총 빈도수를 추출한다 backdata 로 저장된 분서 중에서 범주 c_i 에 속한 문서 전체에 대한 키워드 k_i 의 빈도수를 $freq(c_i,\ k_i)$ 라 하고, backdata 로 저장된 전체 문서에 대한 k_i 의 빈도수를 $freq(k_i)$ 라하면 다음과 같은 관계가 성립한다.

$$freq(k_i) = \sum_{i=1}^{m} freq(c_i,\ k_i)$$

특징 추출을 위한 키워드를 선택하는 기준으로써 FI(Feature Index)를 계산할 수 있다 만일 문서를 나타내는 T 차원의 특징 벡터를 추출하고자 할 경우 FI 가 최대인 T 개의 키워드가 특징 벡터를 구성하는 속성이 된다 키워드 k_i 에 대한 FI 를 $FI(k_i)$ 라 하면 다음과 같은 공식에 의해 구해진다

$$FI(k_i) = \max_{i=1}^{m}\left[\frac{freq(c_i,\ k_i)}{freq(k_i)}\right]$$

위의 공식에 의하면 특정 부문의 범주에 키워드의 빈도수가 집중될수록 FI 의 값이 증가하게 된다 그리하여 FI 가 최대인 T 개의 키워드가 선택되고 문서에 대해 선택된 키워드의 빈도수가 특징 벡터로 형성된다. 이러한 특징 벡터로 전환하여 k-Nearest Neighbor 또는 신경망에 문서의 입력 형태로 사용된다 특징 벡터의 적용 과정은 다음절에서 설명하겠다.

3. 실험 및 결과

이 절에서는 1998 년 5 월 10 일에서 1998 년 5 월 15 까지의 조선 일보 신문 기사 200 개를 학습 데이터로 선정하고 1998 년 5 월 19 일부터 5 월 20 일까지의 신문 기사 80 개를 테스트 데이터로 선정하였다 기사의 범주로는 정치, 경제, 스포츠, 정보통신 4 개의 범주로 선정하고, 학습 데이터의 200 개 기사에는 각 범주당 50 개, 테스트 데이터의 80 개 기사에는 각 범주당 20 개로 하였다

문서의 특징 추출 과정에 있어서는 2 절에서 언급한 과정에 의해 20, 30, 40, 50 차원 4 가지로하여 성능 테스트를 하였다.

각 차원의 벡터에 대한 키워드의 문서에 대한 빈도수를 특징 벡터로 나타내고 Nearest Neighbor 알고리즘 또는 신경망의 모델인 역전파를 적용한 결과가 표 1 과 같다.

표 1 Nearest Neighbor 알고리즘과 신경망에 의한 테스트 결과

	20	30	40	50
1-Nearest Neighbor	0.3125	0.3375	0.3750	0.4625
5-Nearest Neighbor	0.3375	0.2500	0.2875	0.5000
10-Nearest Neighbor	0.3250	0.2625	0.3250	0.3625
신경망 (역전파)	0.3500	0.3000	0.5250	0.5375

위의 표 1 에서 제시된 수치 데이터는 전체 문서 수에 대해 정답의 범주가 부여된 문서 수의 비율을 나타낸다 Nearest Neighbor 알고리즘인 경우, 50 차원의 특징 벡터에 대해서 테스트 데이터와 가장 근접한 5 개의 학습 데이터를 선택할 경우 최고의 성능을 나타내었으며, 신경망을 이용할 경우 50 차원의 벡터를 사용할 경우 최고의 성능을 나타내었다. 그리고 신문 기사 범주화에 대해여 통계적 방법인 Nearest Neighbor 보다 신경망이 향상된 성능을 나타내었다

4. 결론

문서 범주화에 대한 한 예로서 신문 기사 분류에 대해 신경망과 Nearest Neighbor 알고리즘의 비교를 제시하였다.

그리하여 4 절의 실험 결과에 의해 신경망의 접근인 Nearest Neighbor 알고리즘보다 우수하다는 것을 알 수 있다.

문서 범주화인 경우 특징 벡터로 표현하기 위해 상당 수의 키워드가 필요하다. 동일 범주의 문서라도 사용된 가워드는 공통 키워드보다 상이한 키워드가 다수이기 때문에 이를 포괄 적으로 포함하기 위해 그 키워드 개수는 많아진다. 문서 범주화에 있어서 수천 차원의 특징 벡터를 형성해야 하며 이를 위한 학습 데이터에 사용될 문서도 상당히 필요하게 된다.

4 절에서 제시된 문서 범주 결과에 의하면 상대적으로 신경망의 성능이 우수하지만 정답률이 60%미만이기 때문에 논문에서 접근한 방법으로는 실제 상황에 이용될 수 없다. 이의 성능을 향상하기 위한 알고리즘의 개선이 필요하다

이를 개선하는 방법으로써 각 범주 당 1 개의 신경망이 할당되고, 범주를 판정하는 대신 범주에 대하여 긍정, 부정의 이분법으로 답하여 해당 범주의 신경망에는 긍정의 정답을 비 해당 범주의 신경망에는 부정의 정답을 제시하도록 하여 이를 개선하는 것이 앞으로의 과제이다.

5. 참고 문헌

[1] M S. Chen, J. Han, P.S. Yu. "Data Mining: An Overview from a Database Perspective", pp866-883, IEEE Transaction on Knowledge and Data Engineering Vol 8 6, 1996.

[2] K M Decker and S. Forcardi, "Technology Overview: A Report on Data Mining", Technical Report CSCS TR-95-02, Swiss Scientific Computing Center, 1995.

[3] N J. Radcliffe and P D. Surry, "Co-operation through Hierarchical Competition in Genetic Data Mining", Technical Report EPCC-TR-94-09, Edinbyrgh Parallel Computing Center, 1994.

[4] H. Lu. R. Setiono, and H. Liu, "Effective Data Mining using Neural Network", pp957-961, IEEE Transaction on Knowledge and Data Engineering Vol 8:6, 1996

[5] V I Frants, J Shapiro. and V G. Voiskunskii. Automated Information Retrieval: Theory and Methods, Academic Press, 1997.

[6] M A Hearst, "Text Data Mining: Issues Techniques and the Relation to Information Access", http://www sims.berkeley edu/~ hearst/talks/dm-talk, 1997.

[7] J Eldredge, "Text Data Mining an Overview", http://www.cs.columbia.edu/~radev/cs6998/class/ cs6998-09-02/img001htm, 1997

[8] A D Marwick, "Mining on Text Data", http //www.software ibm.com/data/iminer/fortext/ presentations/marwick/index htm, 1998.

[9] R Seiffert, "Text Mining and Retrieval: A Development View", http //www software.ibm.com/data/ iminer/fortext/presentations/seiffert/index htm, 1998

[10] D.D. Lewis, "Representation and learning in Information Retrieval", Dissertation of PhD. the Graduate School of the University of Massachusetts, 1992.

[11] M.T. Hagan, H B Demuth, and M. Beale, *Neural Network Design*, PWS Publishing Company, 1995.

[12] J A Freeman and D M Skapura, Neural Networks Algorithms, Application and Programming Techniques, Addison-Wesley Publishing Company, 1992.

[13] 이성환, *패턴 인식의 원리 I*, 홍릉 과학 출판사, 1994.

[14] T. Joachims, "Text Categorization with Support Vector Machines: learning with Many Relevant Features", LS-8 Report 23, Technical Report in University of Dortmund, 1997.

[15] D. Koller and M. Sahami. "Hierarchically Classifying Documents using very few Words", pp170-178, The proceedings of International Conference on Machine learning, 1997

[16] M. Sahami. M. Hearst, and E Saund, "Applying the Multiple Case Mixture Model to Text Categorization", appeared. the proceedings of International confernce on Machine Learning, 1996

[17] L.S Larkey and W.B Croft, "Combining Classifiers in Text Categorization", pp289-297 The Proceedings of SIGIR 96, 1996

[18] D D. Lewis. R E.Schapire. J.P.Callan, and R.Papka, "Training Algorithm for Linear Text Classifier", pp298-315, The Proceedings of SIGIR 96, 1996

[19] 이성환. *패턴 인식의 원리 II*, 홍릉 과학 출판사, 1994.