전자우편 문서의 자동분류를 위한 다중 분류기 결합
(Combining Multiple Classifiers for Automatic Classification of Email Documents)

이 지 행 † 조 성 배 ††
(Jee-Haeng Lee) (Sung-Bae Cho)

요 약 디지털 형태의 문서가 널리 퍼지고 끊임없이 증가함에 따라 이를 자동으로 가공하고 처리하는 문서 자동분류의 중요성이 널리 인식되고 있다. 최근의 문서 자동분류는 k-최근접 이웃, 결정트리, Support Vector Machine, 신경망 등의 다양한 기계학습 기법을 이용하여 연구되고 있다. 그러나 많은 연구가 잘 조직된 데이타 집합을 이용하여 연구결과를 보여주고 있으며, 실제 문제에의 응용성에는 큰 비중을 두지 않고 있다. 본 논문에서는 문서분류의 응용시스템인 질의 자동응답시스템에 적용할 수 있는 다중 분류기 결합 방법을 제안하고 실제 전자우편 문서의 분류문제를 해결한다. 첫째로, 다중신경망을 이용한 문서분류를 제안한다. 제안한 방법은 최대값 결합, 신경망 결합을 통해 성능의 향상을 가져온다. 둘째로, 여러 분류기의 결합을 통해 문서분류의 성능을 개선한다. 본 논문에서는 투표 결합방법, Borda 결합, 신경망 결합방법 등을 적용하여 여러 분류기의 결합을 수행하였다. 실용 가능성을 분석한 실험결과 90%이상의 정확율을 보여 제안한 방법이 실용적일 수 있음을 알 수 있었다.

키워드 : 문서 자동분류, 다중방법결합, 다중신경망, 자동질의응답

Abstract Automated text classification is considered as an important method to manage and process a huge amount of documents in digital forms that are widespread and continuously increasing. Recently, text classification has been addressed with machine learning technologies such as k-nearest neighbor, decision tree, support vector machine and neural networks. However, only few investigations in text classification are studied on real problems but on well-organized text corpus, and do not show their usefulness. This paper proposes and analyzes text classification methods for a real application, email document classification task. First, we propose a combining method of multiple neural networks that improves the performance through the combinations with maximum and neural networks. Second, we present another strategy of combining multiple machine learning classifiers. Voting, Borda count and neural networks improve the overall classification performance. Experimental results show the usefulness of the proposed methods for a real application domain, yielding more than 90% precision rates.

Key workds : document classification, multiple methods combination, multiple neural networks, automatic query response

1. 서 론

문서의 자동분류는 1960년대 정보검색의 한 분야로 연구되기 시작하였다. 1980년대 말까지는 주로 이론적인 연구에 머물러 있었으며, 실제 응용시스템도 전문가가 수작업을 통해 생성해낸 규칙을 기반으로한 방법을 통해 주로 구현되었다[1]. 그러나 1990년대에 접어들어 컴퓨터가 널리 보급되고, 인터넷이 발전함에 따라 디지털형태의 정보가 급격히 증가하기 시작하여 정보의 과잉현상이 나타나게 되었다. 따라서 많은 양의 정보를 자동으로 가공하여 분류하는 문서 자동분류 분야의 중요성이 널리 인식되기 시작하였으며, 현재에 이르기까지 다양한 이론과 방법들이 깊이있게 연구되고 있다[1, 2, 3].

최근 문서 자동분류 분야의 중요한 특징은 다양한 기계학습 기법을 기반으로 연구가 이루어지고 있다는 점

· 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRP-2000-005-C00012).
† 비 회 원 : (주)다음소프트 자연어처리연구소 연구원
easygo@daumsoft.com
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@cs.yonsei.ac.kr
논문접수 : 2001년 8월 3일
심사완료 : 2001년 11월 20일

이다[1, 3, 4, 5]. 많은 양의 정보가 다양한 분야에서 끊임없이 등장하고 있으므로, 많은 수작업을 필요로 하는 도메인 지식을 이용한 지식기반 시스템으로 처리해 내기에는 한계가 있다. 따라서, 대량의 정보를 자동으로 분류하는 기술이 필요하게 되어, 패턴인식 등 인공지능 분야에서 많이 연구되고 있는 다양한 기계학습 기법들이 문서 자동분류 분야에 적용되고 있다.

이러한 문서 자동분류를 이용하여 새로운 뉴스를 분류하거나, 회사로 들어오는 방대한 양의 전자우편을 해당 부서로 라우팅 시키는 등 실제 응용에 관한 연구가 수행되고 있다[4, 8]. 하지만, 대부분의 경우 문서 자동분류의 실용성보다는 각 분류기의 학습 성능에 초점을 맞춘 연구이기 때문에 실제 문제에 직접 적용하기에는 많은 어려움이 있다. 즉, 많은 문서 자동분류의 실험들이 Reuter 문서집합이나 TREC 문서집합 등 실험을 위해 잘 조직된 데이타 집합을 이용하기 때문이다[1]. 실제 데이타의 경우 많은 잡음, 극도의 클래스간 데이타 개수 불균형, 전체적인 데이타 부족 등 많은 학습상의 어려움 때문에 잘 조직된 데이타 집합과는 다른 학습 양상을 보여준다. 또한, 문서분류의 실험결과로 정확율, 재현율 등의 수치만을 나열하고 실제 응용을 위한 학습 방법이나 실험결과를 분석한 연구는 찾아보기 힘들다.

본 논문에서는 문서 자동분류를 실제 문제에 적용하기 위한 기계학습 방법으로 다음의 두 가지를 제안한다.

첫째로, 다중신경망을 이용한 문서 자동분류를 제안한다. 여러 패턴인식 문제에서 하나의 큰 신경망을 사용하는 것 보다 작은 문제를 푸는 부 신경망의 결합을 통해 좋은 성능을 낼 수 있다고 알려져 있다[6, 7, 8, 9]. 특히, 잡음이 많거나 클래스가 불균형한 데이타의 학습을 위해서는 이러한 접근방식이 효과적이다. 본 논문에서는 하나의 클래스를 하나의 신경망을 이용하여 모델링하며, 이러한 여러개의 신경망을 적절한 방법으로 결합하는 방법을 제안한다.

둘째로, 여러 분류기의 결합을 통해 문서분류의 성능을 개선하는 방법론을 제안한다. 분류를 위한 기계학습 알고리즘들은 각각의 특성에 따라 다른 분류 양상을 보인다. 따라서 분류 결과를 효과적으로 결합할 경우 전체적인 분류 성능의 향상을 가져올 수 있다. 본 논문에서는 다수결 결합방법, Borda 결합방법, 신경망 결합방법을 적용하여 여러 분류기의 결합을 수행한다.

이러한 기계학습 방법론을 (주)다음커뮤니케이션의 포탈사이트인 한메일넷의 사용자 질의 문서집합을 이용하여 검증하였다. 사용자 질의 문서집합의 자동분류 결과는 질의 자동응답시스템에 응용되어, 운영자가 직접 답변하는 것에 비하여 사용자는 빠른 응답을 받을 수 있으며, 운영자의 작업량을 크게 줄일 수 있다는 장점이 있다. 하지만 질의 자동응답시스템에 적용되기 위해서는 분류의 신뢰도를 극대화할 필요가 있다. 이를 위하여 본 논문에서는 패턴인식 분야에서 주로 사용되는 인식율, 오인식율, 기각율을 통한 분석과 정보검색 분야에서 사용되는 정확율, 재현율, F-measure 등을 종합적으로 이용하여 분석한다. 특히, 정확율에 중요도를 부여하는 F-measure를 사용하여 분류기의 학습성능을 분석하고 응용성을 논한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구배경인 문서분류의 정의 및 방법론과 문제점을 살펴보고, 널리 사용되고 있는 기계학습을 위한 문서분류기에 대하여 살펴본다. 3장에서는 다중신경망을 이용한 문서분류 시스템과 학습방법, 결합방법을 제안하며, 4장에서는 여러 분류기의 결합을 통한 문서분류 방법론을 제안한다. 5장에서는 실험결과의 분석을 통하여 제안한 방법의 유용성을 입증하고, 6장에서 결론을 맺는다.

2. 배 경

문서 자동분류는 새로운 문서를 미리 정의된 클래스로 대응시키는 일련의 작업을 말한다. 미리 정의된 클래스의 집합을 $C=\{c_1, c_2, \ldots, c_n\}$, 새로운 문서의 집합을 $D=\{d_1, d_2, \ldots, d_m\}$이라 할 때, 문서의 자동분류는 알려지지 않은 분류함수 $g: C \times D \rightarrow \{0,1\}$에 근접한 함수 $f: C \times D \rightarrow \{0,1\}$를 만들어 내는 것이다. 여기서 분류함수 g는 임의의 클래스 c_i $(1 \le i \le n)$에 대한 임의의 문서 d_j $(1 \le j \le m)$의 멤버쉽함수이다. 즉, $f(c_i, d_j)$가 1일 경우 d_j가 c_i에 속하는 것을 의미하며, 0일 경우 그렇지 않음을 의미한다. 따라서 f는 분류함수 g에 가능한 한 근접하게 만들어야 하며, 근접한 정도가 f의 성능, 즉 자동분류 성능의 기준이 된다.

자동분류의 정의에서 살펴볼 수 있는 것처럼 하나의 문서는 하나 이상의 클래스에 대응될 수 있으며, 어떠한 클래스에 대응되지 않을 수도 있다. 분류함수에 의해 결정된 문서의 클래스개수에 따라 단일분류 문제와 다중분류 문제가 존재한다. 단일분류 문제는 분류함수 g가 모든 문서에 대하여 오직 하나의 클래스에 대해서만 1의 값을 가지는 경우이며, 다중분류 문제는 g가 임의의 문서 d가 속하는 클래스의 개수에 제한이 없는 (0과 m 사이의 임의의 값) 경우이다. 본 논문에서는 단일분류 문제를 위한 문서분류기에 관하여 연구한다.

2.1 문서분류 시스템

문서분류기를 만드는 것은 분류함수를 자동으로 생성

해낼 수 있는 기계를 만드는 것을 의미한다. 이러한 문서분류기를 만들기 위하여 1960년대 이래로 많은 연구가 이루어져 왔으나, 최근에는 각종 기계학습 기법을 이용하여 도메인지식에 독립적이고, 대량의 문서를 다룰 수 있으며, 사람의 수작업이 적게 들어갈 수 있는 방법론이 주를 이루고 있다[1, 3]. 왜냐하면 기계학습은 미리 분류된 데이타를 학습하여 자동으로 분류기를 생성해 낼 수 있으므로, 충분한 학습 데이타만 존재한다면 원하는 성능의 분류기를 얻을 수 있기 때문이다.

일반적으로 문서분류 시스템의 작동은 다음과 같다. 텍스트 형태의 문서는 전처리기를 통하여 수치벡터 형태의 데이타로 변환된다. 이 수치벡터는 분류기에 효과적으로 적용되기 위하여 속성선별 모듈을 거친다. 이렇게 얻어진 데이타를 이용하여 분류기가 적절한 분류를 위한 학습을 수행한다. 최종적으로 분류과정을 통해 실제 시스템에 적용하기 위한 최적의 성능을 얻어낸다.

2.2 질의메일 분석

사용자 질의 자동분류 문제는 문서 자동분류의 하나로 간주될 수 있다. 사용자들의 질의들을 수집하여 분류한 후 학습과정을 거치면, 새로운 사용자의 질의를 미리 정의된 하나 또는 하나 이상의 클래스로 대응시킬 수 있다. 이와같은 시스템은 관련 답변들과 담당자가 처리해야할 질의 등으로 분류하여, 사용자가 즉각적인 응답을 받을 수 있고 시스템 운영의 효율을 높일 수 있다는 장점이 있다.

한메일넷은 (주)다음커뮤니케이션에서 제공하는 포탈 시스템의 이름이다. 2001년 7월 현재 약 2000만명의 사용자가 이용하고 있어서, 운영자가 일일이 사용자들의 이용관련 질의에 답변하기에는 많은 어려움이 있다.

표 1 빈도수에 따른 한메일넷 질의 분포

클래스 특성	클래스 개수	데이타 개수
빈도가 많은 질의	6	1002 (44.9%)
개별응답 질의	7	585 (26.2%)
통계적 처리가 힘든 질의	36	127 (5.7%)
기타	18	518 (23.2%)
계	67	2232 (100.0%)

본 논문의 실험을 위하여 약 한달간 한메일넷 사용자의 실제 질의들의 표본을 수집하였으며, 표 1은 질의 집합의 특징을 간략히 보여준다. 빈도가 많은 몇 개의 클래스에 질의가 편중되고 있으며, 많은 클래스는 실제 데이타의 수가 작아 통계적으로 처리가 힘들다. 이러한 한메일넷 사용자 질의 자동분류 문제를 통해 살펴볼 수 있는 특징은 다음과 같다.

- 정확율 문제 : 자동응답 시스템에 적용하기 위한 문서분류는 재현율보다는 정확율이 중요한 지표로 작용한다.
- 문서집합의 클래스별 불균형 문제 : 일반적으로 FAQ 메일의 분류문제는 특정 클래스에 사용자 질의 문서가 편중되는 경향을 나타내어 학습이 어렵다[3, 10, 11].
- 클래스개수 문제 : 분류해야할 개수가 많으면 기계학습기법을 직접 적용하기에 어렵다[6, 8, 11].
- 속성선별 문제 : 모든 색인어를 입력으로 사용할 경우 학습 성능이 저하되기 때문에 적절한 속성을 추출해 내어야 한다[5, 12].
- 기각 문제 : 어떤 클래스의 문서 개수가 불충분할 경우와 새로운 문서가 대응될 클래스가 존재하지 않는 경우에 기각이 필요하다.

3. 다중신경망을 이용한 문서분류

이 장에서는 다중신경망을 이용한 문서분류 방법을 제안한다. 우선 전처리, 속성선별 과정을 포함한 전체 시스템 구성을 설명한다. 그리고 나서, 실제 분류기로 사용될 다중신경망 방법의 학습과 최대값 결합, 신경망 결합, 진화연산 결합 등의 방법을 제안한다.

3.1 전처리

전처리는 사용자의 질의를 색인어 집합으로 추출하는 과정과 이를 수치 벡터로 변환하는 과정으로 나누어 볼 수 있다. 일반적인 문서분류 시스템에서 색인어 집합 추출은 영문 문서의 경우 영어단어 사전을 가지고 있는 형태소 분석기를, 한글 문서의 경우 한글단어 사전을 가지고 있는 형태소 분석기를 사용한다. 하지만 본 논문에서 사용한 한메일넷 질의집합은 범용적인 한글 형태소 분석기를 이용하여 분석할 경우 문서분류기의 학습을 위한 색인어를 효과적으로 추출하지 못한다. 그것은 전문가가 작성한 뉴스기사나 논문과는 달리 일반 사용자들이 작성한 문서를 대상으로 하므로, 맞춤법이 틀린 표현이 많으며, 통신상의 속어나 약어 등을 많이 사용하여 색인어 추출이 어렵기 때문이다. 이를 해결하기 위하여 실제 문서집합에 대한 속성선별, 동의어 처리, 비속어 처리, 띄어쓰기 처리 등이 고려된 형태소 분석기에 대한 깊이 있는 연구가 필요하다.

본 논문에서는 사용자의 질의를 분석하여 색인어의 사전을 생성하고, 새로운 입력 문서에 대해 대응되는 단어를 추출하는 방법을 사용한다. 이때 한메일넷 질의의 분류에 중요한 단서가 될 수 있는 부사나 형용사 등도 추가된다. 또한, 통신상의 속어나 약어, 동의어에 대한 사전을 유지하여 정규화 시키며, 맞춤법에 맞지 않더라

도 문서분류에 중요한 색인어일 경우 추출해 내는 과정을 수행한다.

색인어 집합으로 표현된 문서를 벡터공간 모델(Vector Space Model)을 이용해 수치벡터 형태의 값으로 변환한다[1, 2]. 이는 추출된 색인어가 문서에 나타나는 빈도수와 색인어가 나타난 전체 문서의 개수를 기반으로 계산되는 $tf-idf$ 식에 의하여 구해진다. 문서 d_i의 j번째 키워드 w_{ij}의 가중치는 다음과 같이 표현된다.

$$w_{ij} = tf_{ij}\log(N/df_j) \quad (1)$$

여기서, tf_{ij}는 j번째 색인어의 문서 i에서의 빈도수, df_j는 키워드 j가 전체 문서집합에서 나타나는 문서의 개수, N은 총 문서의 개수를 의미한다. 이와 같은 방법으로 문서 d_i는 벡터 $(w_{1i}, w_{2i}, \ldots, w_{Ki})$로 표현된다. 여기서 K는 총 색인어의 개수를 의미한다.

3.2 속성선별

사전에 등록된 모든 색인어를 분류에 이용할 경우 벡터의 차원이 너무 커지게 되어 패턴인식의 성능 및 속도의 저하를 가져온다[5]. 본 논문에서는 색인어 선택 방식인 χ^2-statistic 방법을 이용하여 각 클래스별로 중요한 색인어를 추출하고 분류에 잡음으로 작용하는 색인어를 제거하였다.

χ^2-statistic 방법은 일반적으로 계산량이 적고, 성능이 우수한 속성선별 방법으로 문서분류 시스템에 널리 사용되고 있다[12]. χ^2값은 색인어와 클래스의 독립성을 계산하는 것으로 수치가 높을수록 키워드가 해당 클래스의 분류에 중요함을 나타낸다. 그 식은 다음과 같다.

$$\chi^2(t,c) = \frac{N(N_{r+}N_{n-} - N_{r-}N_{n+})^2}{(N_{r+}+N_{r-})(N_{n+}+N_{n-})(N_{r+}+N_{n+})(N_{r-}+N_{n-})} \quad (2)$$

여기서, N은 문서의 개수이며 r과 n은 문서가 클래스내에 속하는지 아닌지를, +와 -는 그 색인어가 문서내에 존재하는지 아닌지를 나타낸다.

이렇게 추출된 클래스 중요도는 분류기의 특성에 따라 다른 방식으로 적용된다. Support Vector Machine과 본 논문에서 구현한 신경망처럼 각 클래스별로 이진 결정을 내리는 분류기를 생성하여 결합하는 방식을 사용하기 위해서는 식(2)에서 구한 χ^2값을 직접 이용하여 상위 값을 가지는 색인어들을 속성으로 이용한다. k-최근접 이웃, 결정트리, 단일신경망과 같이 모든 클래스에 대하여 한가지의 속성선별만 수행하는 경우는 각 키워드의 전체 클래스에 대한 중요도를 계산하여야 한다. 이를 위하여 χ^2_{avg} 또는 $\chi^2_{\max}$중 하나를 이용하며 그 식은 다음과 같다[12].

$$\chi^2_{avg}(t) = \sum_{i=1}^{n} P_r(c_i)\chi^2(t,c_i) \quad (3)$$

$$\chi^2_{\max}(t) = \max\{\chi^2(t,c_1), \chi^2(t,c_2), \ldots, \chi^2(t,c_n)\} \quad (4)$$

여기서 $P_r(c_i)$는 임의의 문서가 c_i에 속할 확률을 의미하며, 본 논문에서는 $\chi^2_{\max}$를 이용하여 속성선별을 수행하였다.

3.3 다중신경망 문서분류

속성선별 과정을 통하여 얻어진 데이타를 이용하여 학습을 수행한다. 본 논문에서 제안하는 다중신경망 방법은 단일신경망이 분류해야할 클래스 하나를 모델링하고[13], 이러한 신경망들 전체의 결과를 종합하여 최종 결론을 도출한다. 그림 1은 다중신경망 문서분류기 시스템을 나타낸다.

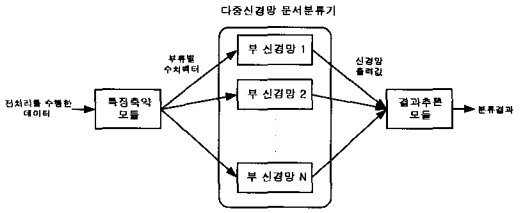

그림 1 다중신경망 문서분류기

이러한 방법의 가장 큰 장점은 문제분할(task decomposition)에 있다[11]. 복잡한 문제를 작은 부문제로 나누어 학습을 용이하게 할 수 있으며, 적절한 결합을 통해 최적의 결과를 도출해 낼 수 있다. 특히, 본 논문에서 적용한 사용자질의 문서집합과 같이 클래스개수가 많을 경우 단일신경망으로 분류를 수행하려면 큰 어려움이 따른다. 반면, 다중신경망 방법은 하나의 클래스를 작은 부신경망으로 모델링하므로, 전체 클래스의 개수가 늘어나더라도 큰 어려움 없이 학습을 수행할 수 있다.

또한, 사용자질의 문서집합은 클래스별 개수 불균형 문제를 가지고 있다. 이러한 문서집합을 단일신경망으로 학습할 경우 공정하지 못한 학습을 유도할 수 있으며, 전체적인 성능 저하를 가져올 수 있다[8]. 다중신경망 방법은 이러한 성능 저하에 영향을 받지 않으며, 각 클래스별 복잡도에 따른 신경망을 구성할 수 있으므로 이의 효과적인 해결책을 제시해 준다.

마지막으로, 속성선별 과정을 통해 각 클래스에 최적화된 속성을 따로 추출할 수 있으므로 클래스별 학습의 성능향상을 가져올 수 있다. 이를 통해 각 부신경망은 작은 입력벡터를 가지고 좋은 성능을 얻을 수 있다. 따

라서 전체적인 성능의 개선과 학습속도 향상을 가져올 수 있다.

3.4 다중신경망 학습

다중신경망은 각 부분류기가 해당 클래스를 최적으로 모델링하면서 동시에 전체적인 분류기능을 최적화 하도록 학습되어야 한다. 이러한 학습은 최대상호정보(Maximum Mutual Information, MMI)를 통해 유도될 수 있다[14, 15]. 클래스 c에 대한 상호정보는 간략하게 다음과 같이 정의된다.

$$I_c = \log P(D_d|\lambda_c) - \log \sum_{x=1}^{N} P(D_d|\lambda_x) \quad (5)$$

여기서, D_c는 클래스 c에 속하는 문서 집합을, N은 전체 클래스의 수를, λ_c는 클래스 c의 모델을 의미하며, $P(D_d|\lambda_x)$는 클래스 c에 속하는 문서가 클래스 x의 모델에 속할 확률을 의미한다. 결국 클래스 c를 위한 학습은 식(5)의 상호정보 I_c를 극대화할 수 있도록 모델의 집합 $\lambda=\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$이 최적화 되어야 한다. 전체 클래스를 위한 상호정보는 다음과 같다.

$$I = \sum_{c=1}^{N} \left\{ \log P(D_d|\lambda_c) - \log \sum_{x=1}^{N} P(D_d|\lambda_x) \right\} \quad (6)$$

결국 모델집합 λ는 식(6)의 상호정보 I를 최대화하도록 선택되어져야 한다. 다중신경망 문제에서 각 클래스별 모델은 부신경망이 되므로 전체 부신경망의 집합이 상호정보 I를 최대화하도록 학습되어야 한다.

그런데, 여기서 상호정보 I를 최대화하기 위해서는 양의 값을 가지는 $\log P(D_d|\lambda_c)$를 최대화하고 음의 값을 가지는 $\log \sum_{x=1}^{N} P(D_d|\lambda_x)$를 최소화하여야 한다. 결국 전체적인 다중신경망의 학습은 각 부신경망이 해당 클래스의 문서에 대해서는 최대의 소속확률을 가지고, 클래스에 속하지 않는 문서에 대해서는 최소의 소속 확률을 가지도록 학습되어야 한다[15].

이러한 조건을 만족시키기 위하여, 본 논문에서는 각 부신경망 λ_c의 학습시에 해당클래스에 속하는 문서집합 D_c와 그렇지 않은 문서집합 $\overline{D}_c$로 나누어, 하나의 출력값을 가지는 부 신경망 λ_c는 D_c에 속하는 문서는 1의 출력값을, $\overline{D}_c$에 속하는 문서는 0의 출력값을 내도록 학습하였다.

3.5 다중신경망의 결과추론

위와같이 학습된 부 신경망의 출력값을 결합하여 최종결과를 추론한다. 이때 각 부신경망의 출력값을 CSV(Classification Status Value)로 사용하여 결과추론을 수행할 수 있다. 다중신경망(Multiple Neural Networks, MNN)이 생성하는 클래스 c_i에대한 문서 d의 CSV는 $CSV_i^{MNN}(d)$로 표현한다.

(1) 최대값 결합

결합방법중 가장 간단한 것으로, 각 부신경망의 출력값 중 최대의 값을 가지는 신경망을 선택하는 방법이다. 문서 d에 대한 다중신경망의 분류 결과 $f_{MNN}(d)$는 다음과 같다.

$$f_{MNN}(d) = \max_{c_i \in C} \{CSV_i^{MNN}(d)\} \quad (7)$$

그러나, 최대값을 가질 때의 CSV값이 정해진 역치 τ를 넘지 못할 경우 문서분류 신뢰도를 높이기 위해 기각을 수행할 수 있다. 본 논문에서는 역치 τ의 변화에 따른 다중신경망 분류기의 성능을 분석한다.

(2) 다중신경망 결합

신경망의 학습능력을 이용하여 결과추론을 최적화하고자 하는 방법이다. 신경망 결합방식을 사용할 경우, 결과적으로 계층적인 신경망이 생성된다.

결과추론을 위한 신경망의 입력은 각 부신경망의 출력값들로 이루어진 벡터인 $(CSV_1^{MNN}(d), CSV_2^{MNN}(d), \ldots, CSV_N^{MNN}(d))$가 되며, 출력은 클래스개수인 N개로 구성하여 각 노드는 각 클래스에 대응된다.

신경망 결합의 장점은 각 부신경망의 분류성향이 종합적으로 고려되어 최종결과를 추론할 수 있다는 점이다. 즉, 하나의 클래스에 대한 결과를 추론할 때 클래스에 해당하는 분류결과 뿐만 아니라 다른 클래스의 분류결과를 이용하므로 결과적으로 성능의 향상을 가져올 수 있다. 예를들어, 비교적 작은 크기의 출력을 내는 부신경망도 추론과정을 통해 선택될 수 있으며, 최대값을 내는 부신경망도 다른 부신경망의 분류결과에 의해 기각될 수 있다. 이렇게 두단계의 신경망을 거쳐 얻어진 최종 분류결과 또한 신뢰도를 높이기 위해 역치를 이용해 기각을 수행한다.

4. 분류기 결합을 통한 문서분류

이 장에서는 여러 분류기의 결합을 통해 문서분류의 성능을 개선하는 방법에 대하여 설명한다. 문서분류를 위한 기계학습 알고리즘들은 각각의 특성에 따라 다른

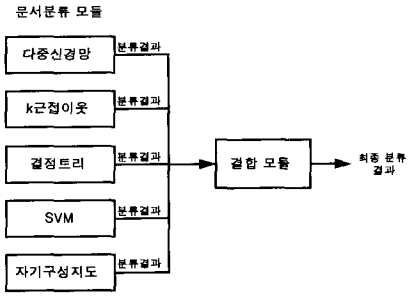

그림 2 여러 분류기의 결합

분류양상을 가진다. 따라서 분류결과를 효과적으로 결합할 경우 전체적인 분류성능의 향상을 가져올 수 있다. 그림 2는 전체적인 결합시스템을 보여준다.

일반적으로 여러 분류기의 결과를 결합하는 방법은 결합에 사용되는 정보의 정도에 따라 3가지 레벨로 나뉠 수 있다[16].

표 2 다중 분류기의 결합 방법

레벨	이용하는 정보	결합방법
추상레벨	각 분류기가 출력하는 하나의 분류결과	다수결, BKS, 베이시안
순위레벨	각 분류기의 모든 클래스별 분류 순위	Borda 함수, Condorect 함수
측정치레벨	각 분류기의 모든 클래스에 대한 분류신뢰도	신뢰값 합산, 신경망

본 논문에서는 다중신경망 분류기와 함께 k-근접이웃[1, 3], 결정트리[5, 17], SVM[18, 19], 자기구성지도[20, 21] 분류기의 결과를 각 레벨의 대표적인 방법인 다수결 결합, Borda 함수, 신경망 결합 방식을 이용하여 최종결과의 향상을 꾀하였다.

(1) 투표결합

추상레벨 결합인 투표결합 방식은 단순하면서도 결합을 위한 추가학습과정이 필요하지 않다. 또한, 결합에 사용될 각 분류기의 성능이 우수할 경우 좋은 결과를 도출할 수 있다. 본 논문에서는 다음의 두가지 투표결합 방식을 이용하였다.

첫 번째 방법은 다수결결합이다. 대부분의 투표결합방식이 사용하는 방법으로서 가장 많은 인식기가 답으로 낸 클래스를 최종 클래스로 선택한다. 만약 최다득표 클래스가 두 개이상 존재할 경우 기각을 수행한다.

두 번째 방법은 만장일치 방식이다. 일반적으로 사용되는 방법은 아니지만, 본 논문과 같이 분류의 목표가 기각율을 높이더라도 분류의 신뢰도를 극대화해야할 경우 유용한 방법이 될 수 있다. 모든 분류기가 만장일치로 분류를 수행하지 않는 경우 기각을 수행한다.

(2) Borda 함수 결합

Borda 함수는 순위레벨 결합방식의 대표적인 방법이다. 클래스 c_i에 대하여 인식기 k가 출력한 순위 r^i_k에 따라 $(N- r^i_k)$를 구하고 이 값들을 합산하여 가장 점수가 큰 클래스가 결과로 결정되는 방식이다. Borda 함수는 다음과 같다.

$$F_{borda}(f_k(d)) = \max_{c_i \in C}(B_i(d)) \quad (8)$$

$$B_i(d) = \sum_{k=1}^{K}(N- r^i_k(d)) \quad (9)$$

(3) 신경망 결합

가중치레벨 결합방식인 신경망 결합방식은 각 분류기의 클래스별 신뢰도를 이용하여 최종결과를 위한 학습을 수행하는 방법이다. 결과결합에 사용되는 신경망은 분류기의 개수가 K, 분류해야할 클래스의 개수가 N일 때 $K\times N$개의 입력노드와 N개의 출력노드를 가진다. 신경망의 입력벡터 $Input_{NN}$는 다음과 같이 정의된다.

$$Input_{NN} = CVS^1 \times CVS^2 \times \ldots \times CVS^K \quad (10)$$

단, $CVS^k = (CVS^k_1, CVS^k_2, \ldots, CVS^k_N)$ 이다.

5. 실험결과

5.1 실험환경

본 논문에서 제안한 문서분류기의 성능검증은 약 한 달간 수집된 한메일넷 사용자 질의 표본 2204개를 대상으로 이루어 졌으며, 질의집합 중 임의로 선택하여 1718개의 문서를 학습데이타로, 463개의 문서를 성능 평가를 위한 테스트데이타로 사용하였다. 또한, 분류기 결합을 위해 학습이 필요한 경우 학습데이타 중 10%의 데이타를 임의로 추출하여 검증데이타로 사용하였다.

질의의 특성상 직접 분류가 필요없이 운영자에게 전달되어야할 클래스가 존재한다. 개별답장이 필요한 질의 클래스, 답장이 필요없는 질의클래스, 질의의 빈도가 현저히 떨어지는 클래스 등은 운영자에게 전달하여 직접 처리할 클래스로 간주하였다. 이 클래스 또한 하나의 '기각 클래스'로 간주하여 학습에 참여시켰는데, 그 분포는 표 3과 같다.

표 3 응답종류에 따른 분류

	클래스 개수	질의 개수
응답되어야할 질의	18	1475 (66.9%)
운영자에게 전달해야 할 질의	46	729 (33.1%)

다중신경망 학습시에 부신경망은 각 클래스별로 구성하였으며, 운영자에게 전달해야 할 클래스집합을 하나의 기각클래스로 하여 부신경망을 구성하였다. 각 부신경망의 은닉노드의 개수는 실험에 의해 결정하였으며, 각 부신경망의 인식율의 향상이 없을 때까지 학습을 수행하였는데, 모든 경우에서 반복회수 200내에 학습이 완료되었다.

5.2 분류기의 성능분석

이 절에서는 다중신경망의 분류성능을 k-최근접 이웃, 결정트리, SVM, 자기구성지도 분류기, 단일 신경망

과 비교분석한다. 다중신경망의 결합은 최대값 결합방식을 이용하였다.

첫째로 분류기의 성능분석을 위해 주로 패턴인식에서 사용하는 지표, 즉 인식율, 오인식율, 기각율 및 신뢰도를 이용하여 수행한다. 여기서, 기각율은 전체문서 중 기각 문서의 비율, 신뢰도는 전체 문서중 오인식하지 않은 문서의 비율을 의미한다.

신뢰도의 조절이 가능한 분류기의 경우 신뢰도를 극대화한 경우와 인식율을 극대화한 경우로 나누어 결과를 구하였다. 표 4에서 제안한 다중신경망이 단일신경망을 포함한 모든 분류기에 비해 좋은 인식율과 신뢰도를 보임을 알 수 있다. 단, 전반적인 분류기들의 결과가 기본적인 패턴인식 실험에 비하여 인식율이 떨어지고 기각율이 높은 것을 알 수 있다. 이는 잘 가공된 데이타집합이 아닌 실제 데이타 집합을 이용했기 때문에 생기는 현상이며, 본 논문에서는 기각율을 높임으로써 신뢰도를 높여 실제 응용가능하도록 하였다.

표 4 패턴인식 지표를 이용한 분류기별 성능

분류기		인식율	오인식율	기각율	신뢰도
다중신경망	인식율극대화	0.741	0.140	0.119	0.860
	신뢰도극대화	0.650	0.052	0.298	0.948
단일신경망	인식율극대화	0.741	0.149	0.110	0.851
	신뢰도극대화	0.659	0.076	0.265	0.924
k-최근접이웃	인식율극대화	0.704	0.145	0.151	0.855
	신뢰도극대화	0.647	0.054	0.309	0.946
결정트리		0.693	0.182	0.126	0.818
SVM	인식율극대화	0.568	0.097	0.335	0.903
	신뢰도극대화	0.544	0.039	0.417	0.961
자기구성지도	인식율극대화	0.518	0.060	0.421	0.940
	신뢰도극대화	0.427	0.002	0.570	0.998

다른 분류기의 분류 성능에 대하여 좀더 살펴보면, k-최근접 이웃 방법이 비교적 좋은 성능을 보였으며, 상대적으로 SVM과 결정트리 분류기의 성능이 조금 떨어짐을 볼 수 있었다. SVM의 경우 인식율은 현저히 떨어지나, 높은 신뢰도를 얻을 수 있었고, 결정트리는 기각율을 조절할 수 있는 알고리즘을 도입하여 더욱 높은 신뢰도를 얻을 수 있는 방법을 찾아야 할 것이다.

둘째로 기각율의 변화에 따른 오류율을 그래프로 나타내어, 분류기의 전체적인 분류성향을 정성적으로 관찰한다. 그림 3은 다중신경망, k-최근접 이웃, SVM, 자기구성지도, 단일신경망 분류기에 대하여 역치의 조작에 의한 기각율 변화에 따른 오류율의 추세를 나타낸다.

기각율 변화에 따른 오류율 그래프는 분류기의 곡선이 아래쪽에 위치할수록 분류의 성능이 우수함을 나타낸다. 그림 3은 다중신경망의 성능이 가장 우수함을 나타내고 있으며 단일 신경망, kNN, SVM, 자기구성지도의 순으로 성능이 우수함을 나타내고 있다.

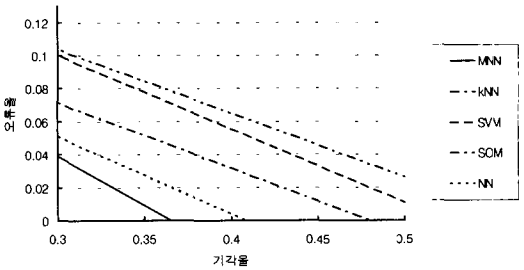

그림 3 분류기별 기각율 변화에 따른 오류율

세째로 정보검색 분야에서 사용하는 지표를 이용해 분류기별 성능을 분석한다. 표 5는 하나의 클래스에 대한 실제 클래스와 분류기의 분류결과가 나타날 수 있는 경우의 개수 도표(Contigency Table)를 보여준다.

표 5 하나의 클래스에 대한 분류결과 경우의 개수

클래스 c_i		분류기의 분류	
		속함	속하지 않음
실제 분류	속함	A_i	B_i
	속하지 않음	C_i	D_i

정확율은 분류기를 통해서 수행된 분류결과 중 정확한 것의 비율, 재현율은 분류되어야할 결과중 정확한 것의 비율을 의미하며, 표 5를 통해 정의되는 정확율과 재현율은 다음과 같다.

$$\text{정확율} = \frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} A_i + \sum_{i=1}^{N} C_i} \tag{11}$$

$$\text{재현율} = \frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} A_i + \sum_{i=1}^{N} B_i} \tag{12}$$

또한, 이렇게 정의된 정확율과 재현율을 결합한 지표로서 F-measure가 있다.

$$F_\beta = \frac{(\beta^2+1)\cdot Pr \cdot Re}{\beta^2 \cdot Pr + Re} \tag{13}$$

여기에서, Pr은 정확율을 Re는 재현율을 의미하며, β는 정확율과 재현율의 중요도를 조절할 수 있는 매개변수다. β가 0일 경우 F_β는 정확율을 나타내며, β가

+∞일 경우 재현율을 나타낸다. 보통의 경우 β를 1로 하여 정확율과 재현율을 동일한 중요도로 간주한다.

표 6 정보검색 지표를 이용한 분류기별 성능

분류기		F_1	$F_{0.3}$	정확율	재현율
다중신경망	F_1최대	0.739	0.775	0.783	0.699
	$F_{0.3}$최대	0.614	0.837	0.903	0.465
단일신경망	F_1최대	0.783	0.761	0.766	0.712
	$F_{0.3}$최대	0.614	0.837	0.903	0.464
k-최근접 이웃	F_1최대	0.692	0.744	0.755	0.639
	$F_{0.3}$최대	0.500	0.835	0.962	0.331
결정트리		0.690	0.697	0.699	0.682
SVM	F_1최대	0.500	0.714	0.780	0.368
	$F_{0.3}$최대	0.437	0.753	0.878	0.288
자기구성지도	F_1최대	0.428	0.674	0.761	0.298
	$F_{0.3}$최대	0.425	0.691	0.786	0.294

표 6은 각 분류기의 성능을 정확율, 재현율, F_1과 $F_{0.3}$이 최대값을 가질 때의 수치를 이용하여 분석한 결과를 보여준다. 일반적으로 정보검색분야의 분석에서는 정확율과 재현율을 동일한 비율로 조합한 F_1이 최대일 경우의 분석에 초점을 맞춘다[3]. 하지만 본 논문에서는 이러한 분석을 포함하여 정확율에 더욱 가중치를 두는 $F_{0.3}$의 결과를 함께 분석하였다. 이는 자동 문서분류의 실용성의 검증을 위한 지표로서 의미를 지니며, 기존의 분석방법과는 다른 학습의 목표를 제시할 수 있다. 분석결과 다중신경망의 분류결과가 가장 우수함을 알 수 있으며, 패턴인식적 분석과 비교할 때 정확율은 신뢰도에, 기각율의 역은 재현율에 대응되어 유사한 지표로 분석됨을 알 수 있다.

또한, 본 논문에서 학습의 지표로 삼고있는 $F_{0.3}$은 분류기의 만족할 만한 재현율에서 정확율을 극대화시킬 수 있음을 알 수 있다. 그림 4에서 볼 수 있듯이 다중신경망의 정확율이 0.9를 넘고, 재현율이 급격히 떨어지기 전에 $F_{0.3}$이 극대값을 가지게 되어 적절한 역치를 결정할 수 있는 지표가 된다. 아울러 이러한 역치 조절을 통해 다중신경망의 전체적인 성능을 향상시킬 수 있음을 알 수 있다.

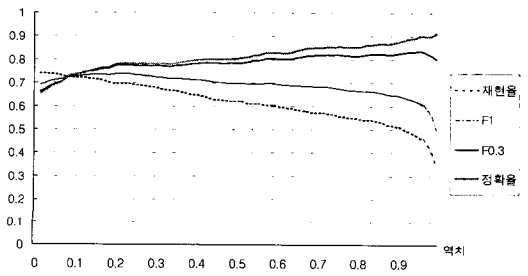

그림 4 역치 변화에 따른 다중신경망의 성능변화

5.3 결합방법의 성능분석

이 절에서는 먼저 제안한 다중신경망 결합방법의 성능을 분석한다. 최대값결합과 신경망결합을 사용할 때, 전체적으로 유사한 결과를 보여주고 있으나 최대값결합에 비하여 신경망결합이 좀더 우수한 결과를 내고 있다. 충분한 크기의 학습데이타를 사용할 경우 분류기의 성능이 좀더 향상 될 수 있을 것이다. 표 7은 정보검색 지표를 이용한 다중신경망 결합방법의 성능을 보여준다.

표 7 다중신경망 결합 성능

결합방법	F_1	$F_{0.3}$	정확율	재현율
최대값결합	0.614	0.837	0.903	0.465
신경망결합	0.649	0.849	0.904	0.507

다음으로 다중분류기의 결합을 통하여 성능을 개선시킨 결과를 알아본다. 최대값결합을 이용한 다중신경망과 k-근접이웃, 결정트리, SVM, 자기구성지도를 결합한 결과는 표 8과 같다.

$F_{0.3}$값이 가장 큰 경우는 다수결 투표결합, 만장일치 투표결합, 신경망 결합방법이었다. 이 방법 모두 결합에 사용된 부분류기보다 성능이 향상됨을 알 수 있었으며, 특히 투표결합방식은 추가적 학습이 필요없는 간단한 알고리즘으로 좋은 결과를 도출해 낼 수 있었다. 그러나 Borda 함수 결합은 분류기의 성능향상에 실패하였는데, 전체 데이타의 크기가 충분하거나 사용한 문서분류기가 더욱 많을 경우 적합할 것으로 생각된다. 결과적으로 여러 분류기를 결합하여 문서분류를 수행할 경우 개별 분류기를 사용하였을 때 보다 더 나은 분류기를 얻을 수 있었다.

표 8 정보검색 지표를 이용한 다중분류기 결합 성능

결합방법	F_1	$F_{0.3}$	정확율	재현율
투표결합(다수결)	0.741	0.845	0.870	0.648
투표결합(만장일치)	0.644	0.847	0.903	0.500
Borda함수 결합	0.648	0.810	0.852	0.523
신경망결합	0.773	0.830	0.843	0.713

6. 결 론

본 논문에서는 문서분류의 실제응용을 위한 분류성능

최적화 방법을 제안하였다. 대상으로 삼은 응용시스템은 사용자 질의 응답시스템으로서 문서분류기의 분류 신뢰도와 정확율을 극대화 해야하는 것이다. 이를 위하여 다음의 두가지 방법을 제안하였다.

첫째로 다중신경망을 이용한 문서분류를 수행하였다. 단일신경망이 아닌 클래스별 부신경망을 구성하여 학습시키는 방법을 제안하였으며, 이의 효과적인 결합을 위해 최대값결합과 신경망 결합을 이용하였다. 또한 기각을 통해 분류기의 정확율을 극대화하는 방법론을 제안하였으며, 실험결과 기존의 분류기에 비하여 좋은 성능의 문서분류기를 얻을 수 있었다.

둘째로 여러 분류기의 결합을 통한 분류성능의 향상을 시도하였다. 각기 다른 분류성향을 나타내는 분류기들의 결과를 종합하여 개선된 성능의 문서분류기를 구축하였다. 본 논문에서는 대표적인 분류기 결합방식인 투표결합, Borda 함수결합, 신경망 결합방법을 이용하여 제안한 다중신경망 방법과 함께, k-최근접 이웃, 결정트리, SVM, 자기구성지도 등의 분류기를 결합하였다. 실험결과 하나의 분류기보다 여러 분류기를 결합하였을 경우 더 나은 성능을 보임을 알 수 있었다.

제안한 분류기를 실제 사용자들의 질의 집합인 한메일넷 사용자 질의 메일 집합에 적용하여 분석하였다. 분류기를 통해 사용자 질의 집합을 자동으로 분류하여 자동 답변 또는 운영자에게 전달하는 실제 질의응답 시스템에 응용하여야 하므로 분류기의 신뢰도가 중요하다. 본 논문에서는 이를 위한 지표를 제안하고 다중 분류기 결합방식으로 학습을 수행하여 최적의 성능을 얻어낼 수 있었으며, 신뢰도 90% 이상의 분류성능으로 실제 시스템에 응용될 수 있음을 알 수 있었다.

추후 연구로는 좀더 큰 크기의 데이타 집합을 이용하여 제안한 분류기의 성능을 객관적으로 검증할 필요가 있다. 이를 위하여 현재 약 5만건의 실제 사용자의 이메일 질의 집합을 수집하여 실험을 준비중에 있다. 또한, 범용적인 형태소 분석기가 아닌 일반인이 작성한 전자우편 문서집합의 자동 분류를 위한 형태소 분석기에 관한 연구가 필요하다. 이를 통해 문서의 속성을 효과적으로 선별해 내어 전체적인 분류기의 향상을 가져올 수 있을 것이다.

참 고 문 헌

[1] F. Sebastiani, "Machine Learning in Automated Text Categorisation," *Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.

[3] Y. Yang and X. Liu, "A Re-examination of Text Categorization Methods," *Proceedings of the 22h Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99)*, pp. 42-49, 1999.

[4] Y. Yang, T. Ault and T. Pierce, "Combining Multiple Learning Strategies for Effective Cross Validation," *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1167-1182, 2000.

[5] S. M. Weiss, et al., "Maximizing Text-Mining Performance," *IEEE Intelligent System*, pp. 63-69, July/August 1999.

[6] T. Caelli, L. Guan, and W. Wen, "Modularity in Neural Computing," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1497-1518, September 1999.

[7] A. J. C. Sharkey, "On Combining Artificial Neural Nets," *Connection Science*, vol. 8, no. 3/4, pp. 299-314, 1996.

[8] R. Anand, et.al., "Efficient Classification for Muticlass Problems Using Modular Neural Networks," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 117-124, 1995.

[9] L. S. Larkey and W. B. Croft, "Combining Classifiers in Text Categorization," *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96)*, pp. 289-297, 1996.

[10] R. Anand, et.al., "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 962-969, 1993.

[11] B. Lu and M. Ito, "Task Decomposition and Module Combination Based on Class Relations: A Modular Neural Network for Pattern Classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1244-1256, September 1999.

[12] Y. Yang and J. P. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *In Jr. D. H. Fisher (Ed.), The 14th International Conference on Machine Learning*, pp. 412-420, Morgan Kaufmann, 1997.

[13] R. P. Lippmann, "An Introduction to Computing with Neural Networks," *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 4, no. 2, pp. 4-22, 1987.

[14] Y. Ephraim and L. R. Rabiner, "On the Relations Between Modeling Approaches for Speech Recognition," IEEE Transactions on Information Theory, vol. 36, no. 2, pp. 372-380, March 1990.

[15] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, February 1989.
[16] L. Xu, A. Krzyzak and C. Y. Suen, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418-435, 1992.
[17] J. R. Quinlan, "Decision Trees and Decision-making," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 20, no. 2, pp. 339-346, March/April 1990.
[18] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.
[19] T. Joachims, "Estimating the Generalization Performance of a SVM Efficiently," *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, 2000.
[20] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, "WEBSOM-Self-Organizing Maps of Document Collections," *Neurocomputing*, vol. 21, pp. 101-117, 1998.
[21] 김현돈, 조성배, "한메일넷 질의 자동응답을 위한 이단계 자기구성 지도," *한국정보과학회·춘계학술발표논문집(B)*, pp.481-484, 대구, April 2000.

이 지 행

1998년 2월 연세대학교 컴퓨터과학과 학사. 2001년 8월 연세대학교 컴퓨터산업시스템공학과 석사. 2001년 2월 ~ 현재 (주) 다음소프트 자연어처리 연구소 연구원. 관심분야는 인공지능, 자연어처리, 정보검색, 인공생명 등

조 성 배

1988년 연세대학교 전산과학과(학사). 1990년 한국과학기술원 전산학과(석사). 1993년 한국과학기술원 전산학과(박사). 1993년 ~ 1995년 일본 ATR 인간정보통신연구소 객원 연구원. 1998년 호주 Univ. of New South Wales 초청연구원. 1995년 ~ 현재 연세대학교 컴퓨터과학과 부교수. 관심분야는 신경망, 패턴인식, 지능정보처리