학습방법개선과 후처리 분석을 이용한 자동문서분류의 성능향상 방법

최 윤 정[†] · 박 승 수[††]

요 약

자동문서분류는 문서의 내용에 기반하여 미리 정의된 항목에 자동으로 할당하는 작업으로서 효율적인 정보관리 및 검색등에 필수적인 작업이다. 기존의 문서분류성능 향상을 위한 연구들은 대부분 분류모델 자체를 개선시키는 데 주력해왔으며 통계적인 방법으로 그범위가 제한되어왔다. 본 연구에서는 자동문서분류의 성능향상을 위해 데이터마이닝 기법과 결합허용방법을 이용하는 개선된 학습알고리즘과 후처리 방법에 의한 RTPost 시스템을 제안한다.

RTPost 시스템은 학습문서 선택작업 이전에 분류항목 설정의 문제를 다루며, 분류함수의 성능보다는 지정방식의 문제점을 감안하여 학습과 분류 후처리 프로세스를 개선하려는 것이다. 이를 통해 분류결과에 중요한 영향을 미쳐왔던 학습문서의 수와 선택방법, 분류모델의 성능등에 의존하지 않는 안정적인 분류가 가능하였고, 이를 분류오류율이 높은 경계선 인접영역에 위치한 문서들에 적용한 결과 높은 정확율을 얻을 수 있었다. 뿐만 아니라, RTPost 프로세스를 진행하는 동안 능동학습방법의 장점을 수용하여 학습효과는 높이며 비용을 감소시킬 수 있는 자가학습방법(self learning)방법의 효과를 기대할 수 있다.

키워드 : 자동 문서분류기법, 능동적학습방법, 자가학습방법, 계층적 분류, 텍스트마이닝, 데이터마이닝, 결합허용기법

Reinforcement Method for Automated Text Classification using Post-processing and Training with Definition Criteria

Yun Jeong Choi[†] · Seung Soo Park[††]

ABSTRACT

Automated text categorization is to classify free text documents into predefined categories automatically and whose main goals is to reduce considerable manual process required to the task. The researches to improving the text categorization performance(efficiency) in recent years, focused on enhancing existing classification models and algorithms itself, but, whose range had been limited by feature based statistical methodology.

In this paper, we propose RTPost system of different style from any traditional method, which takes fault tolerant system approach and data mining strategy. The 2 important parts of RTPost system are reinforcement training and post-processing part. First, the main point of training method deals with the problem of defining category to be classified before selecting training sample documents. And post-processing method deals with the problem of assigning category, not performance of classification algorithms.

In experiments, we applied our system to documents getting low classification accuracy which were laid on a decision boundary nearby. Through the experiments, we shows that our system has high accuracy and stability in actual conditions. It wholly did not depend on some variables which are important influence to classification power such as number of training documents, selection problem and performance of classification algorithms. In addition, we can expect self learning effect which decrease the training cost and increase the training power with employing active learning advantage.

Key Words : Automated Text Categorization(classification), Active Learning, Self Learning, Hierarchical Classification, Text Mining, Data Mining, Fault Detection

1. 서 론

최근 인터넷이 활성화되면서 웹사이트의 사용자들로부터 생성되는 문서들의 규모가 엄청난 속도로 증가하고 있다. 이들 문서들은 그 내용이 점점 복잡하고 다양해지고 있으며 정보로서의 잠재적 가치가 있는 문서가 있는 반면 그렇지 않은 것들도 공존하기 때문에, 이들을 분석하기 위한 문서 분류에 대한 관심이 높아지고 있다. 문서의 분류는 신문 등의 뉴스서비스 뿐만 아니라 연구논문 등의 체계적인 관리를 위해서도 매우 중요하다. 그러나 이러한 문서량의 폭주로 인하여 이들에 대한 수작업 분류가 한계에 도달하게 되었고 문서의 내용에 따른 분류의 자동화에 대한 요구가 급증한

※ 이 논문은 2004년도 두뇌한국21사업에 의하여 지원되었음.
(This work was supported by the Brain Korea 21 Project in 2004.)
† 준 회 원 : 이화여자대학교 컴퓨터학과 박사과정
†† 정 회 원 : 이화여자대학교 컴퓨터학과 교수
논문접수 : 2005년 6월 8일, 심사완료 : 2005년 11월 1일

것이다. 특히 필터링 규칙을 피해가며 대량으로 유포되는 광고성 문서들의 경우만 보더라도 컴퓨터에 의한 문서의 내용파악은 현재로서는 한계가 있기 때문에 자동문서분류의 정확성과 결과의 신뢰도를 높이는 문제가 대두되고 있다.

자동문서분류에 있어서 정확도를 좌우하는 요인으로는 학습(training)방법과 분류모델을 들 수 있다. 학습방법에서는 사이즈가 작으면서 정보량이 큰 문서를 학습 데이터로 선택하는 것이 관건이며, 선택과정에서 전문가의 수작업을 필요로 하기도 한다. 텍스트 문서 분류를 위해 사용되는 분류기는 대부분 벡터모델과 확률모델에 기반한다. 이러한 모델들은 문서들을 'bag-of-words'로 표현하며 이 어휘자질(feature)로 표현된 특성벡터(feature vector)를 사용하여 분류기를 학습하고 이를 카테고리 할당에 그대로 적용하고 있다. 그리고 최종적으로 해당문서를 분류항목에 할당할 때 가장 높은 값을 갖는 항목으로 지정한다.

기존의 문서분류 성능향상을 위한 연구에서는 분류모델 개선방법을 주로 다루고 있는데, 여러 분류기들의 장점을 취하여 상호보완하거나 해당 도메인에 적합한 특성(feature) 가중치 함수를 새로 설계하는 방법을 취하고 있다. 그러나 이러한 방법은 도메인에 대한 사전(dictionary)구성과 선택된 자질(feature selection)의 영향을 많이 받는 문제가 있다[2, 5]. 그리고 복잡한 학습과정을 거쳤더라도 단순한 지정방식에 의해 분류가 결정되어진다는 점으로 볼 때 전체적으로 비효율적인 점이 많다[4, 10].

본 논문에서는 자동문서분류의 성능개선을 위하여 이들과는 다른 접근을 제안한다. 제안방법은 학습방법과 분류지정방법의 두 부분으로 나누어진다. 먼저 학습과정에서는 분류경계(*decision boundary*) 영역에서 많이 발생하는 오류율을 최소화하기 위해 경계 인접부근을 새로운 항목으로 간주하여 분류항목을 정의하고 있다. 이는 문서선택의 문제가 아니라 분류항목 설정의 문제로 변환하는 것이다. 그리고 분류지정방법에서는 분류결과의 후처리 분석을 위한 과정을 정의함으로써 분류함수 성능의 문제가 아니라 지정방식의 문제를 다루고 있다. 제안방법의 타당성 평가를 위한 실험에서는 타분류기와의 성능을 비교하기 위해 다의적인 요소가 많은 문서집단을 대상으로 하였고, 오류문서가 포함된 학습집합에 대해서도 테스트하였다. 각각의 항목마다 정확율(precision, positive predictive power)을 계산한 결과 기존의 방법보다 분류오류율이 현저히 낮아짐을 확인할 수 있었는데, 이는 특성들을 공유하여 분류경계가 뚜렷하지 않은 문서들의 처리가 가능했기 때문으로 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 최근의 문서분류시스템의 성능개선을 위한 연구를 정리하고, 3장에서는 본 연구에서 제안한 개선된 학습방법과 후처리 방법을 설명한다. 4장에서는 이를 토대로 한 실험내용을 보이며5장에서 결론 및 향후 연구에 대해 논의한다.

2. 자동문서분류의 성능개선을 위한 연구

자동문서분류의 정확도와 신뢰도를 높이기 성능향상을 위한 연구 초기에는 문서분류 알고리즘들이 주 대상이 되어 기존의 여러 알고리즘들의 장점을 결합시키는 방법이 대부분이었다. 이후에는 기계학습방법에 대한 중요성이 강조되어 기존의 학습방법을 개선하거나 새로운 학습방법이 제안되기도 하였다. 또한, 이와 다른 접근방법으로 분류결과의 신뢰도를 높이기 위한 연구가 있는데, 학습데이터와 분류기들의 구성방법에 변화를 주는 기법인 앙상블(ensemble)방법이 그 예이다[18]. 앙상블이란 여러 개의 분류기들의 집합으로서, 각각의 분류기를 서로 다른 학습데이터로 훈련을 시키고, 학습된 분류기들의 결과들을 결합하는 방법을 말한다. [9]의 연구에서는 서로 다른 데이터로 훈련된 임시 분류기를 사용하는데, 정답을 알고있는 문서 d_i에 대해 각각의 결과값의 변이(variance)를 계산하여 정답과 가장 가까운 분류기를 선택하고 있다. 최근의 학습과정과 분류알고리즘에서의 연구내용은 다음과 같이 정리할 수 있다.

2.1학습방법(Training Algorithms)

일반적인 학습과정에서는 학습문서집합으로 지정할 문서 개체의 수와 선택, 조합적인 구성의 문제를 다룬다[5]. 여기서 학습효과는 반복횟수와 문서 양이 많을수록 좋은 결과를 얻는 것이 아니라 학습문서들이 지닌 정보량에 의존한다는 점을 지적할 수 있다.

최근의 학습방법들은 학습문서의 선택과 수에 있어서 임의의 선택이 아닌 학습에 도움이 될 가능성이 큰 문서를 신중히 선택하는 것에 초점을 둔다. 이러한 방법을 능동학습(active learning) 또는 질의학습(query learning)이라 하고, 보다 적은 양의 문서와 학습비용으로 높은 효과를 내는 것을 목표로 한다[9, 13, 14]. 학습문서의 선택이 중요한 이유는 적절치 못한 학습문서로 인해 특정 자질이 갖는 예측력이 편향되어질 수 있으며, 전혀 무관한 항목으로 분류하는 오류를 낳을 수 있기 때문이다. 반면, 동일한 학습데이터라 할지라도 각 범주에 해당하는 학습문서집합을 구성하는 방법에 따라서 분류기준과 예측력이 달라질 수 있다. 샘플링알고리즘, 부스팅알고리즘들은 학습문서들의 구성문제를 다루는 대표적인 방법이다. 이들은 대체적으로 임시적으로 사용되는 여러 개의 분류기들을 구성하고 그 결과의 신뢰도를 추정하는데 응용된다[18, 19]. 예를 들면 다음과 같다. 학습문서 구성의 조합을 달리하여 $f = \{ f_1 f_2,\cdots,f_n\}$로 이루어진 분류기들을 생성할 수 있다. 이들 분류기로부터 각기 다른 결과값이 출력되면 이들의 변이값을 이용하여 최적의 분류함수 f_k 를 얻어내는 것이다. 앞서 언급한 앙상블방법은 최적의 분류함수를 만들기 위해 투표(voting) 결합, 가중치(weighting)에 의한 결합, 베이지언 결합, 신경망 결합방법을 이용하고 있다.

전통적인 기계학습모델 중 보다 좋은 결과를 예상할 수 있는 것은 교사(감독)학습(supervised learning) 방법이다. 교사 학습의 훈련방법으로 주로 사용되는 것은 네트워크의 출력값과 목표값간의 차이를 측정하는 것이다. 이를 흔히 오류율이라 하며, 다음 훈련에서 이 오류율의 증감을 측정

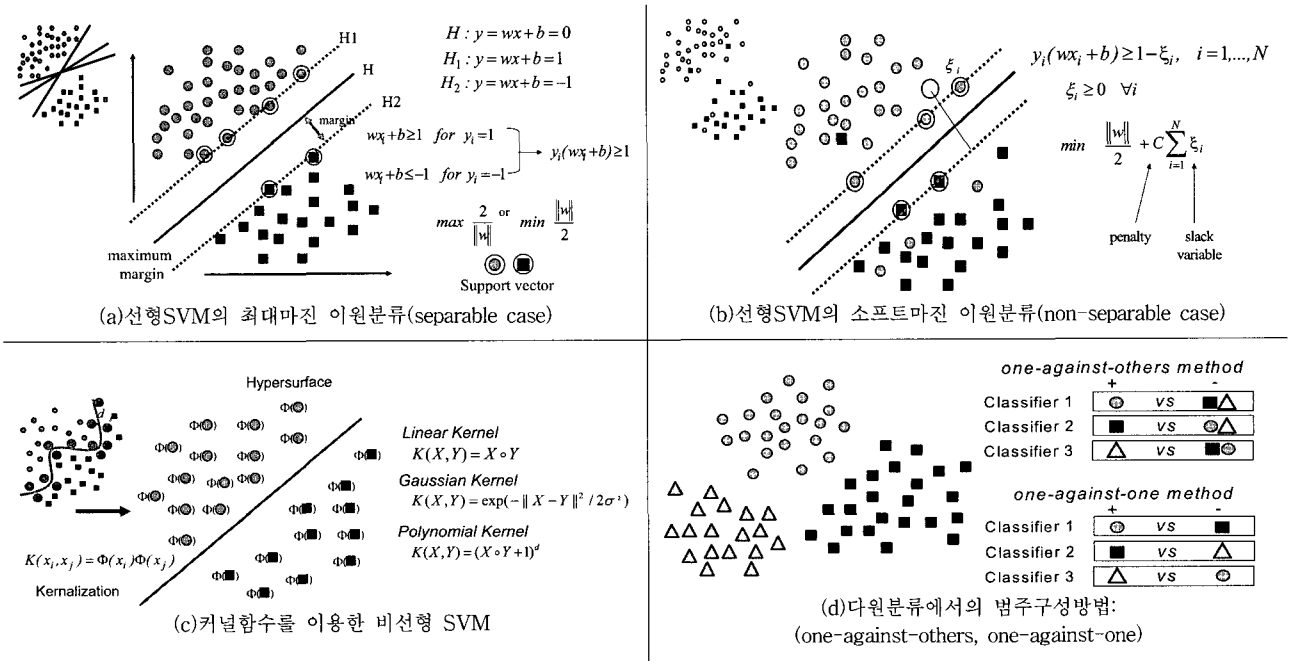
하면서 목표에 점차 접근해 나가는 것이다. 이 알고리즘들은 분류(classification) 모델에 적합하며 피드포워드(feedforward) 파라다임에 해당한다. 교사학습은 기본적으로 입력공간에서 주어지는 데이터에 대해 전문가의 감독하에서 수동적인 학습을 취하는 형태로 이루어지기 때문에 전체 프로세스내에서 전문가 개입에 대한 부담을 가중시킬 수 있다. 위에서 설명한 학습알고리즘들 역시 교사학습에 속하며 적극적인 전문가의 개입이 요구된다. 이러한 이유로 인적비용이 적은 비교사학습(unsupervised learning)을 분류문제에 적용하기도 하는데, 부가적으로 교사학습 지향적인 사항들을 첨가하여 학습비용은 줄이면서 적정수준의 정확도를 보증할 수 있는 방안으로도 연구되고 있다.

2.2 문서 분류알고리즘(Classification Algorithms and Techniques)

문서를 군집화하거나 분류하기 위한 모델로는 기계 학습 분야에서 사용되는 알고리즘들이 사용된다. 이는 크게 규칙기반 모델과 연역적 학습모델, 검색모델로 나뉘어 진다. 규칙기반 모델은 학습 문서들에서 나타나는 범주간의 분류를 위한 규칙을 찾아내거나, 아니면 전문가가 규칙을 지정하여 문서를 분류한다. 연역적 학습모델에는 학습문서에서 자질을 추출하여 확률적으로 접근하는 베이지언 확률모델과[3], 트리구조를 이용하여 자질의 유무로 범주를 결정하는 결정트리모델등이 있다. 또 최근에는 양성/음성자질을 벡터로 표현하고 이들의 차이를 극명하게 하는 지지벡터를 찾는 Support Vector Machines(이하 SVM)등이 각광을 받고있다[11]. 그 외에도 정보검색관점에서 분류할 문서를 질의로 보고 이와 유사한 문서를 찾는 방법인K-Nearest Neighbor(이하 KNN)이 있다. 대체적으로 일반적인 분류문제에서는 이들 중에서 KNN과 선형/비선형 최적화 기법이 가능한 SVM이 가장 좋은 성능을 보이고 있다. 그러나 SVM은 이론상으로 가장 높은 성능을 발휘하지만 현실세계에서는 시공간상의 높은 복잡도로 인해 근사화된 알고리즘으로 구현되기 때문에 이론상의 성능에 미치지 못한다고 평가되고 있다[11, 12]. Naïve Bayesian은 기법이 단순한 것에 비해 결과가 좋고 적용이 쉽다고 알려져 있다[3, 6]. 또한 학습문서집합의 변경 내용에 크게 영향을 받지않는 편이어서 학습문서에 오류가 포함되었을 경우 민감하게 반응하지 않기 때문에 상당히 안정적이라고 보고되고 있다[9, 10]. 이러한 분류기법들이 가진 특성을 중심으로 여러 형태의 변형이 제안되고 있는데, 정확도를 향상시키기 위하여 알고리즘 자체를 개선시키거나 여러 알고리즘을 결합하여 적용하는 등의 형태를 취하고 있다. 본 연구는 경계면에 근접한 문서들의 처리하기 위해 최적의 결정경계면을 찾는 SVM의 방법론을 활용하고 있으므로 SVM의 최적화 방법에 대해 좀더 살펴보기로 한다.

SVM의 최적화 문제는 두 범주를 구분하게 되는 무수히 많은 하이퍼플레인(hyperplain) 중에 어떤 것이 가장 적절한 것인지를 결정하는 것부터 시작한다. (그림 1) (a)에서 w는 단위길이를 갖는 하이퍼플레인과 직교하는 벡터이며, 이때 원점에서 이 하이퍼플레인과의 수직거리는 $|b|/\|w\|$가 된다. 여기서의 최적화 문제는 $max(2/\|w\|)$ 또는 $min(\|w\|/2)$를 찾는 것이고, 라그랑지계수를 도입한 원문제(priminal problem)에 Karuch-Kuhn-Tucker(KKT) 조건과 Wolfe dual problem을 적용시켜 조절변수에 대해 하이퍼플레인의 계수w와b를 결정한다. 그러나 (그림 1) (b)와 같이 선형방법으로는 분리되지 않을 경우가 있다. 이때에는 오분류를 허용하는 것이 불가피한데 이를 위한 최적화문제에 여유변수(slack variable)를 추가하고 있다.

즉, 오분류를 허용하되 하이퍼플레인에서 떨어진 거리에



(a)선형SVM의 최대마진 이원분류(separable case)

(b)선형SVM의 소프트마진 이원분류(non-separable case)

(c)커널함수를 이용한 비선형 SVM

(d)다원분류에서의 범주구성방법:
(one-against-others, one-against-one)

(그림 1) SVM의 최적화방법과 다원분류의 범주구성방법

비례하여 페날티를 준다. 이때, C값이 페날티변수이다. 이 C값에 따라 분류기준이 변하기 마련이지만 결과에 크게 영향을 주지 않는 것으로 알려져 있고, 적절한C값을 찾기 위해서 교차검증(cross-validation)을 활용하여 학습한다.

비선형 SVM에서는 고차원으로 매핑시켜 최적화 시키는데 커널함수를 활용하여 입력벡터 x의 새로운 특징을 추출해낸다. 이는 곧 입력벡터를 고차원 자질 공간에서의 벡터로 변형한 후 선형의 경계선을 찾는 문제로 전환된다. (그림 1) (c)의 대표적 커널함수 이외에도 Sigmoid Kernal, Fisher Kernal, String Kernel등 많은 함수가 제안되고 있다[10, 11, 12].

SVM은 원래 이원패턴 분리를 위한 알고리즘으로 개발되었기 때문에 k-범주의 패턴분리 문제를 위해서는 여러 개의 SVM 이원분류기를 조합하여 다원분류기로 확장을 하게 된다. 이 때 범주는 (그림 1) (d)와 같이 '1대다'(one-against-others)와 '1대1'(one-against-one)로 구성할 수 있다. 전자는 쉽고 단순하지만 분류기의 구분력이 한쪽으로 편향될 수 있다는 특성이 있다. 반면 후자는 너무 많은 분류기가 필요하게 되고 분류기 구축과정에 드는 시간비용이 많이 든다는 단점이 있다. 이와 함께 최적의 다원분류를 위한 분류기 함수조합도 중요하다. 조합방법으로는 k 범주 각각에 대해 이원 결정함수를 구축하여 k개의 범주로 확장하는 승자독식방법(winner-takes-all)이 있고, 이 내용의 단점을 개선하기 위해 제안된 쌍단위 분류방법(pairwise classification)이 있다.

정리한 바와 같이 1979년 Vapnik에 의해 발표된 바 있는 SVM은 최근에 와서야 그 성능을 인정받아 자연어처리분야와 단백질구조예측문제와 같은 여러 분야에 적용되고 있다. 패턴인식분야에서는 SVM을 다원분류기로 확장하는 연구가 활발히 진행되고 있는 반면 문서분류 분야에서는 다원분류기를 중심으로 고찰하는 연구는 적은 편이다. 이는 대량의 문서를 다뤄야 한다는 점과 모든 조합에 대해 학습하고 분류기를 생성하는 비용이 너무 높기 때문이다.

2.3 오분류율이 높은 문서처리를 위한 연구

위와 같이 각종 분류알고리즘들은 가중치를 조절하거나 최적의 경계선을 찾아내는 방법으로 정확성을 개선하려하고 있다. 그러나 복합적인 내용의 문서들은 분류경계 상의 위치가 명확하지 않으며, 이는 그대로 분류결과의 신뢰도에 영향을 준다. 최근의 자질투영법(feature projection)과 분류항목간의 관계(class relevance)에 대한 연구에서는 이러한 문서의 복잡성에 의한 오류를 다루고 있다. 기계학습분야에서 패턴분류를 위해 연구된 자질투영법(feature projection)은 학습데이터를 각 자질의 투영으로 표현하며, 각 데이터에 대해서 투영된 자질들로부터의 투표(voting)에 근거하여 분류가 이뤄진다. 이 방법은 최근 문서분류에 적용하기 위해 새로운 형태의 알고리즘으로 개발되고 있다[7, 8].

$$\text{Relation}(c_i, c_j) = \text{similarity of } (D_i, D_j) \quad (1)$$

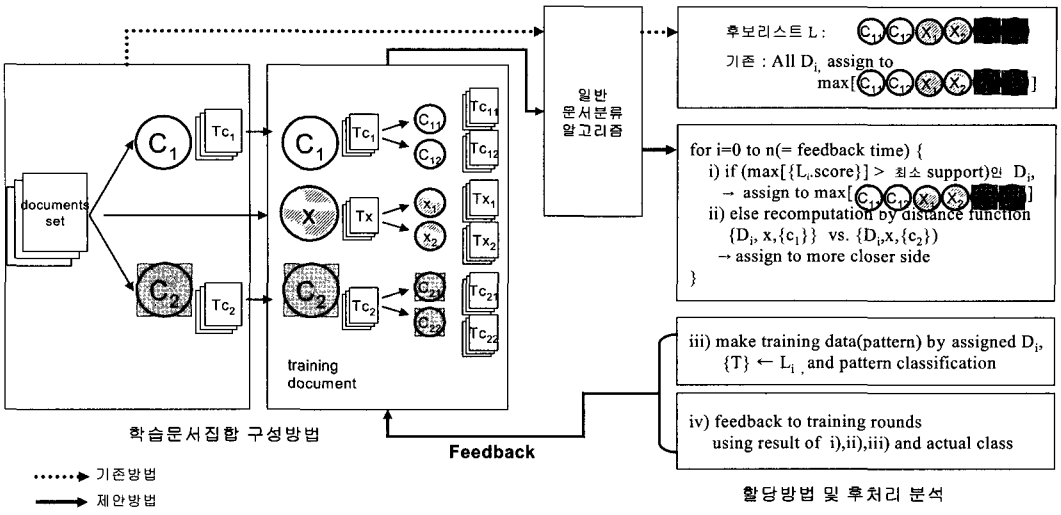
D_n : a set of classified document to target category c_n

그러나 이 방법 역시 용어의 빈도수만을 가지고 문서를 모델링하고 있기 때문에 다양한 내용을 내포하고 있는 문서집합에는 적합하지 않은 면이 있다. 자질의 빈도수를 기초로 한 문서 모델은 어떤 요소와 함께 인접해서 사용되었느냐에 따라 달라지는 자질의 의미를 정확히 반영하기가 어렵기 때문이다. 이러한 문제를 위해서는 시소러스 같은 언어모델을 이용하기도 하며, 분류항목간의 관계(class relevance)를 고려하기도 한다[7]. 식(1)은 관계성을 계산하는 간단한 예로서 학습문서들간의 유사도를 이용하고 있다. 이 방법은 어떠한 문서가 특정 항목과 유사한 관계에 있다면 그 항목과 가까운 관계에 있는 다른 범주들과 상대적으로 높은 유사성을 갖고 있을 것이라는 가정에서 출발한다. 이는 몇몇 문제에서 효과적일 수 있음을 보였으나 문서들간의 유사도와 문서-항목간의 관련정도를 나타내는 수치정보를 단순히 재계산하고 있다는 점은 유사한 범주간의 근소한 차이를 극복하기에는 부족하다. 따라서 유사한 범주간의 구별능력을 극대화 시킬 수 있도록 하는 것을 현안으로 삼고 있다.

3. 학습방법과 후처리분석을 강화한 문서분류 방법: RTPost System

앞 절에서 정리했듯이 다양한 접근방법과 시도들에 의해 자동문서분류의 정확도 향상을 위한 연구가 진행되고 있다. 그러나 분류를 결정짓는 경계영역에 근접한 문서들의 오분류 문제를 개선하는 데에는 별 도움이 되지 못하고 있다. 여기에는 기존의 단순한 지정방식이 갖는 문제점 또한 무시할 수 없다. 본 연구에서는 경계영역에 위치한 문서의 오분류율을 감소시킴으로써 정확도를 향상시킬 수 있는 방법을 제안하고 있다. 제안방법은 학습문서집합을 구성하는 방법과 분류 후 최종항목으로 지정하는 방법의 두 분야로 나뉘며, 이에 따라 학습과 후처리 프로세스를 설계하였다. 또한 제안방법에 의한 시스템은 데이터 마이닝프로세스에 기반하여 설계되었다. 일부 데이터마이닝에서는 신경망이나 유전자알고리즘 같은 특정기법들에만 초점을 두고 있으나 다양한 지식탐사를 위한 개념적인 정보추출의 방법론이자 일련의 과정(process)으로 이해해야 한다는 점을 강조한 것이다. 어떤 문제를 다루는데 정해진 기법이나 규칙이 정해져 있는 것이 아니라 데이터에 따라 혹은 다루어야 할 문제의 성격에 따라 다양한 기법들이 적용될 수 있어야 하기 때문이다. 데이터마이닝을 통해 얻어진 정보는 평가를 통해 다시 마이닝 초기단계에 반영되고 재분석이 되면서 얻게될 결과의 신뢰성을 높여가게 된다. 따라서 지침이 되는 가이드라인이 제시되어야 하며 마이닝된 결과를 비교/평가하고 어떻게 활용할 것인가를 판단하는 인적요소의 역할 또한 중요하다[1].

(그림 2)는 제안방법의 기본개념을 도식화하여 보여주고 있다. RTPost 시스템의 주요 핵심은 학습문서집합의 구성방법과 분류알고리즘을 적용한 이후의 후처리 부분에 잘 나타나 있다. 일반적으로 학습문서집합은 최종분류항목의 집합과 같기 마련이지만 본 연구에서는 경계영역에 대해 다른 시각으로 접근하여 이를 학습문서집합에 추가하고 있다. 이



(그림 2) RTPost(Reinforcement Training+Postprocessing) 시스템의 흐름도

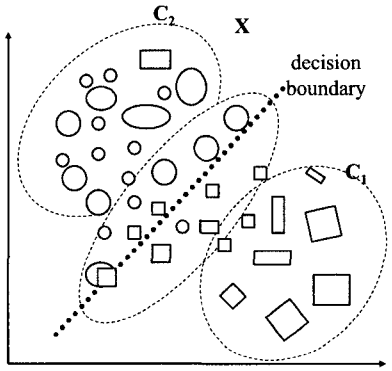
때 분류항목을 계층적으로 세부화하여 앙상블로 구성한다. 앙상블방법은 전문가 한명의 판단보다 여러 전문가에 의견을 종합하여 판단한 결과의 신뢰도가 더 높다는 가정에서 출발한 방법이다. 따라서 앙상블로 구성한 학습문서집합의 구성은 문서분류모델 자체보다는 범주결정방식에 영향을 주고 결과의 신뢰도 향상에 도움이 된다. 후처리 부분은 결함을 예측하고 실행하는 피드포워드(feedforward)방식을 응용한 것이다.

강화된 후처리분석을 통해 성능에 영향을 미치는 자질값과 변수(parameter) 등의 여러 세부적인 요인들에 비교적 둔감하고 효율적으로 동작하도록 한다. 이는 학습문서의 구성방식과 분류알고리즘 성능에 전적으로 의존하지 않는 전체적으로 안정적인 분류 프로세스가 되게한다. 이 프로세스를 진행하는 동안 학습효과는 높이며 비용을 감소시킬 수 있는 자가학습방법의 효과와 분류경계선 인접영역의 오류허용을 감소시키는 효과를 기대할 수 있다.

3.1 학습방법 : 목표항목 설정 및 정의

지금까지의 학습알고리즘은 최종분류항목이 결정된 상태로 수집된 제한된 학습문서 내에서 선택과 구성의 문제를 주시해왔으나, 본 연구에서는 목표항목 정의의 문제로 확장하기로 한다. 단, 제안방법에서는 선택 문제는 다루지 않기로 한다.

(그림 3)의 '○'와 '□'들은 목표항목 C_2, C_1에 해당하는 문서 개체들을 도식화 한 것으로, 그 모양과 크기는 문서내용의 복잡도와 다양성을 표현하고 있다. 경계선 부근에서 겹쳐서 나타나고 있는 문서들은 쓰인 단어의 빈도로만 보면 분류기준이 애매하므로 확실히 어떠한 범주에 할당되는지 판단하기 어려운 성질을 지닌다. 대표적인 예로는 평범한 문서로 가장한 스팸성 문서들이 이에 해당한다. 이러한 문서들의 분류문제에 일반적으로 최종 할당을 위한 목표항목으로만 학습문서를 구성하면 만족할 만한 결과를 기대하기 힘들 것이다. 따라서 RTPost 시스템의 학습방법은 분류경계



(그림 3) 내용의 불확실성이 높아 분류경계에 인접한 문서들의 복잡도

영역을 경계항목으로 지정하여 이에 적합한 학습문서를 두어 구성하도록 제안하고, 최종항목의 결정에서는 이 경계항목과 목표항목간의 거리계산을 통해 보다 가까운 쪽으로 정하는 것이다.

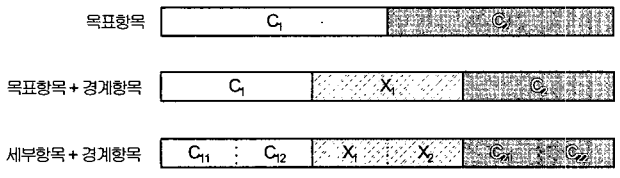
다음은 본 논문에서 사용하는 의미들에 대한 정의 및 소개이다.

- **정의 1 : 목표항목집합(target category set)** 목표항목집합 C = {c_1, c_2,.., c_n}는 궁극적으로 분류하고자하는 최종목표항목 c_i들의 집합을 의미한다. 여기에서 c_i와 c_j는 ($i \neq j$)는disjoint 하다.
- **정의 2 : 세부항목집합(subcategory set)** 목표항목 c_i의 세부항목집합 $c_n' = \{ c_{n1}, c_{n2}, \cdots, c_{nk} \}$ 는 목표항목을 세부적으로 분류한 항목집합으로서 각 c_{nj}는 disjoint하다.
- **정의 3 : 경계항목집합(intermediate category set)** 경계항목집합 $X = \{ x_1, x_2 \cdots x_{n-1} \}$는 목표항목 사이의 경계선 인접부근에 위치한 항목들의 집합으로, 목표항목이 가지는 의미상의 연관성을 고려하여 정의한다. 단, 1차분류시 목표항목으로 분류되지 못한 문서들, 즉 미지정문서도 'X'로 표시한다. 이는 추후 목표항목으로 지정될 개체들이 임시로

지정되는 항목을 의미한다. 여기서 X 의 원소들은 C의 원소들과 disjoint 하지 않을 수 있다.

- **정의 4: 후보항목리스트(list of candidate category)** 후보항목리스트 $L_i = [l_{i1}, l_{i2}, \cdots, l_{im}]$은 입력문서 D_i에 대한 분류결과에 ranking을 부여한 후, 이를 순서대로 열거한 항목의 리스트를 의미한다. 이 때, l_{i1}는 D_i의 최상위후보항목이고, L_i^1이라고도 표기한다.
  여기에서 l_{ij}는 (c, s)의 순서쌍으로, $c \in C \cup X$ 이고, s는 ranking시스템이 부여한 score로서 실수 0과 1사이의 실수이다. 단, m은 순위의 pruning parameter 값으로 목표항목 개수인 n보다 크게 지정되어야 한다.
- **정의 5: 피보트항목(pivot category)** 피보트항목 P는 D_i의 후보리스트 L_i내에서 순위가 가장 높은 경계항목을 말한다. 단, 경계항목들이 인접되었을 경우에는 이들을 병합시킨다.
- **정의 6: 학습문서집합(training documents)** 학습문서집합 $Tr(c_i)$은 c_i를 위한 training set을 의미한다.
- **정의 7: 전체학습문서집합** 입력문서 D에 대한 전체학습문서집합 $T = Tr(c_i) \cup Tr(x_j)$ 으로서, $Tr(c_i)$ 는 $\cup_k Tr(c_{ik})$ 이다. 여기서 학습문서집합의 구성내용은 해당문서집합의 분류목표를 반영하고 있다.
- **정의 5: 최상위후보항목** 최상위후보항목 L_i^1는 후보항목리스트 L_i에서 최상위 후보인 l_{i1}을 의미한다.

위에서 경계항목 x_1은 c_1과 c_2사이의 항목을 말하며, (그림 3)과 같이 이 두 카테고리에 속하는 문서들의 유사도가 높아 분류경계가 모호한 부분에 정의하도록 한다. 이 경계항목에 속한 개체들은 분류가 진행됨에 따라 적절한 목표항목에 배정된다. 위에서 정의한 내용을 기초로 하여 전체 학습문서집합 T를 구성하게 되는데, 분석목표와 분류의도가 반영되도록 정의한다. (그림 4)는 경계항목과 세부항목으로 목표항목을 구성한 학습문서집합 구성방법의 예를 보이고 있다.



초기 학습문서 집합T에 대하여 이원분류일 경우와 다원분류일 때의 구성방법 :

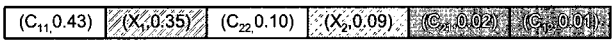
- n= 2 인 이원분류일 때,
  i) $T = \{Tr(c_1), Tr(x), Tr(c_2)\}$: 세부항목을 정의하지 않을 경우
  ii) $T = \{Tr(c_{11}), Tr(c_{12}), Tr(x_{11}), Tr(x_{12}), Tr(c_{21}), Tr(c_{22})\}$: 균등분할로 세부항목을 정의할 경우
- n =3 인 다원분류일 때,
  i) $T = \{Tr(c_1), Tr(x_1), Tr(c_2), Tr(x_2), Tr(c_3)\}$: 세부항목을 정의하지 않을 경우
  ii) $T = \{Tr(c_{11}), Tr(c_{12}), Tr(x_1), Tr(c_{21}), Tr(c_{22}), Tr(x_2), Tr(c_{31}), Tr(c_{32})\}$: 비균등분할로 세부항목을 정의할 경우

(그림 4) 경계항목과 세부항목으로 정의한 RTPost 시스템의 학습문서 집합구성방법

3.2 후처리분석

RTPost 에서 제안하는 후처리방법의 기본 아이디어는 기존 분류체계가 갖는 한계점을 마이닝 관점으로 접근해가면서 분석하고 그 결과를 다시 학습과정에 반영하는 것에 초점을 둔다. 이는 전통적인 범주 할당 방식이 야기하는 문제점을 고려하고 있다. 분석목표 따라 정의한 경계항목과 세부항목으로 학습을 수행하고, 분류알고리즘을 적용하여 전체 후보항목리스트 L을 얻는다. 문서 D_i에 대한 후보항목리스트인 L_i는 분류가능한 범주와 점수치(score)를 쌍으로 하는 순위 리스트이다. <그림5>와 같은 후보항목리스트의 점수치는 적용한 분류모델에 따라 상대적 혹은 절대수치 값으로 얻어진다. 결국 문서 D_i의 후보항목리스트 L_i는 문서와 범주들간의 유사도(similarity)를 나타내는 것이다. 범주 할당작업은 분류수행 결과인 범주와 범주 별 점수치 쌍인 랭킹정보(ranking list)를 분석하는 것에서부터 시작한다.

Document D_i 의 후보항목리스트 L_i



(그림 5) 후보항목리스트(candidate category list)의 예

최근의 개선된 연구에서도 분류지정 방식이 입력문서 D_i에 대해 최상위 항목인 L_i^1 항목으로 할당하는 단순수치와 단순판단에 의존하고 있다. 이는 범주별 수치가 같을 경우 항목 지정 선택의 문제를 다루지 않고 있으며 범주간 근소한 차이를 보일 경우에도 무조건 최상위 항목에 지정하고 있다. 따라서 이 과정의 근본적인 개선이 없이는 정확도 향상에 별 도움을 줄 수 없다고 판단된다. 무엇보다 일반화된 수치분석이 아닌 절대적 수치에 의한 기존의 방법은 상대적인 수치격차나 각 항목의 의미를 전혀 고려치 않기 때문에 분류결과의 신뢰도가 낮다. 나름대로 신뢰도로 허용할 수 있는 수치를 정해두는 것도 최소한의 한 방법이 된다. 이를테면, 절대적 신뢰수치가 현저하게 낮은 문서들은 내용의 정보가치가 떨어진다는 의미를 가짐으로써 우선적으로 가려내는 것이다. 그러나 이러한 방법으로 분류성능을 개선하는 데에는 한계가 따르기 마련이다. 그러므로 본 연구에서는 분류 결정방법에 있어 후보리스트 L_i의 순위별 수치와 항목 모두 분석대상으로 삼고 있으며, 이는 분명 무조건 L_i^1 항목으로 결정짓는 것과는 분명한 차이를 보인다.

본 연구에서 제안하는 후처리 분석은 초기 수치값으로 분석하는 문서할당 단계(step1~step2)와 피드백을 위해 규칙을 생성하는 단계(step3~step4)로 이루어져 있다. <표 1>은 Step1과 Step2의 주요 의사코드이고, 각 과정별 주요내용은 아래와 같다.

〈표 1〉 후보항목리스트의 수치분석에 의한 문서할당규칙

Input : Document D_i, Candidate category list L_i normalized and resorted by descend order

Step1 : for i= 0 to N(= number of input documents) { # 모든 입력문서 D_i에 대해
If (D_i^{size} ≥ min_support) && ((L_i^1score ≥ min_value) || (L_i^1score - L_i^2score ≥ diff_value)) then
assign D_i to L_i^1 # 조건을 만족하는 경우만1순위항목으로 지정
else
assign D_i to X # 미분류항목 X 으로 mark
}

Step2 : for n=0 to N (= number of unassigned documents in step1) { # step1에서X로 mark된 문서 D_i에 대해
for n=0 to N (= number of target category) {
Calculate distance of category between P, c_{nk}

$$Dist(P, c_n) = \sum RD(P, c_{nk}) * w_m$$

}
assign D_i to more closer side c_n
}

3.2.1 step 1 : 후보항목리스트 L_i의 수치값(rank score)과 문서 크기(D_i^{size})에 의한 분석

이 과정은 초기의 분류결과를 1차분류로 하여, 적정값 이상의 점수를 받은 문서들에 대하여 충분히 변별력 있는 값을 가진 문서들에 대해서만 결과를 인정한다. 문서 D_i 에 대한 후보리스트 L_i을 정규화하고 이 수치값들을 이용하여 분류기준을 만든다. 문서 크기 D_i^{size}는 D_i를 이루는 단어의 양(vocabulary size)으로 계산하여 일정기준의 값을 만족하는 문서에 대해서만 1차분류결과를 인정한다. 이 값은 연관규칙 알고리즘에서의 지지도(support)의 의미에 해당하며 사용자가 정하는 파라미터이다. 1위와 2위의 격차가 일정기준 이상을 만족하는 경우에만 1순위 항목으로 지정한다. 이는 연관규칙에서 신뢰도(confidence)의 의미에 해당한다고 볼 수 있다. 위에서 min_support, min_value, diff_value는 모두 사용자가 정의하는 파라미터이며 절단값(cutoff value)의 의미를 갖는다. 이러한 각 일정 수치들은 전체 후보항목리스트 L의 분포값과 격차에 따라 조정하는 것이 바람직하다.

3.2.2 step 2 : 피보트항목 P와 목표항목간의 거리(distance) 계산에 의한 분석

step 1에서의 X로 분류된 미지정문서를 대상으로 후보항목리스트 L_i의 항목패턴을 감안하여 분석하는 과정이다. 이 때, L_i에서 가장 높은 순위의 경계항목인 X_n 즉, 피보트항목 P와 각 목표항목 C_n과의 순위차(distance)를 계산하게 된다. 이 거리계산을 위해 다음의 거리함수를 정의한다. 단, $RD(P, c_{nk})$는 P와 c_{nk}의 순위격차를 의미하며 인접해 있을 경우는 1이다.

$$Dist(P, c_n) = \sum RD(P, c_{nk}) * w_m \quad (2)$$

$c_{nk} \in L_i$: list of candidate category
m= rank order of c_{nk}

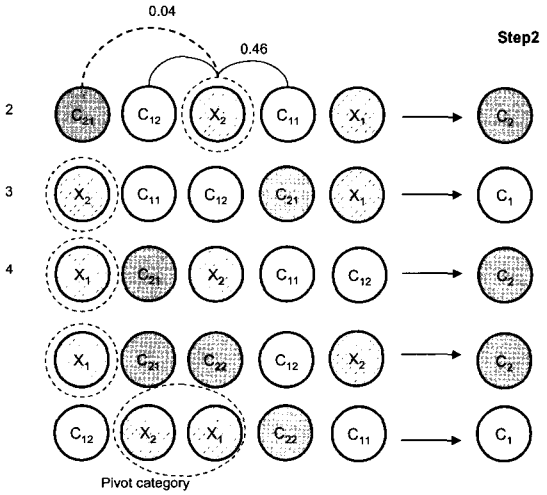
식 (2)에서 같은 그룹의 세부항목끼리는 합하여 계산하며, 로그함수를 이용한 가중치 함수 w_m을 사용하는 이유는 같은 순위격차를 가질 경우에 고순위의 의미를 부여하기 위함이다.

$$w_m = log(\sqrt{(m + \alpha)}) : \text{weight in rank of } c_{nk} \quad (3)$$

a : control parameter

$$\text{Assign: } \min_n \{Dist(P, c_n)\} \quad (4)$$

L_i / D_i	1 ($w_m \approx 0.02$)		2 ($w_m = 0.15$)		3 ($w_m = 0.25$)		4 ($w_m = 0.31$)		5 ($w_m \approx 0.35$)		D_i^{size}	Step1	Step2	Assign	Actual Class
1	C_{21}	.98	C_{11}	.01	X_1	.01	X_2	.01	C_{12}	.00	726.33	C_{21}→C_2	-	C_2	C_2
2	C_{21}	.39	C_{12}	.20	X_2	.17	C_{11}	.13	X_1	.10	31.6	X	C_2	C_2	C_2
3	X_2	.29	C_{11}	.28	C_{12}	.17	C_{21}	.15	X_1	.01	514.42	X	C_1	C_1	C_1
4	X_1	.28	C_{21}	.23	X_2	.17	C_{11}	.16	C_{12}	.15	287.12	X	C_2	C_2	C_2



(그림 6) 후보항목리스트(candidate category list)의 점수치 분석과 피보트항목(pivot category)-목표항목간의 거리분석에 의한 지정 예

〈표 2〉 RTPost 시스템의 유효성(effectiveness) 평가 요소와 예

location 입력문서	단계별 분류 결과값(feedback time =1)						해석(I) 및 지침(A)	
	C_n^1.step1	C_n^1.step2	C_n^2	C_{n+1}^1.step1	C_{n+1}^1.step2	C_{n+1}^2	평가	위치별 예상오류 및 원인
d_1	X	1	1	1	1	1	Good	a) C_n^1.step1 : 학습문서집합과 파라미터 수치
d_2	X	0	1	1	1	1	Good	I: 학습문서집합에 의한 학습력이 낮다.
d_3	X	0	0	1	1	1	Poor	A1:경계항목,목표항목의 문서를 보강하여 학습력을 강화시킨다.
d_4	1	1	1	1	1	1	Fair	A2:cut-off value의 범위을 조정한다.
d_5	1	0	1	1	1	1	Fair	b) C_n^1.step2 : 목표항목간의 거리계산
d_6	1	1	0	1	1	1	Poor	I:문서의 복잡도가 높으므로 경계항목과 목표항목간의 거리에 차이가 없다.
d_7	1	0	0	1	1	1	Poor	A:분류항목을 세부화하여 학습문서집합을 재구성
d_8	0	1	1	1	1	1	Good	c) C_n^2(step3): 학습패턴
d_9	0	0	1	1	1	1	Good	I: 수치 및 거리에 근거한 학습패턴의 구분력이 낮다.
d_{10}	0	1	0	1	1	1	Poor	A:사례를 수집하여 학습패턴의 학습력 강화한다.
	√	√	√	√				

where C_n^e.process, n=0,1,2 ⋯ feedback time, e : 분류입력형태, 1=documents 2=candidate lists of documents, process=step1, step2

식 (3)에서 $_m$은 항목의 절대 순위값이며 조절상수 a를 사용하여 순위별 가중치 값을 조정할 수 있다. 이와 같이 순위가 높을수록 작은 값을 가지는 w_m은 식 (2)에 반영되어 같은 격차라도 현저한 차이를 부여할 뿐 만 아니라 관계의 긴밀성을 나타낸다. (그림 6)은 피보트항목과 목표항목간의 거리를 근거로 한 지정방식을 보이고 있다.

3.2.3 step 3 : 후보항목패턴에 의한 분류규칙 생성

위의 진행과정을 통해 미지정문서에 대한 분류가 점차적으로 이루어지며, 단순히 위의 두 단계를 거치는 것만으로도 기존 방법에서의 오류가 어느 정도는 해결될 수 있음을 알 수 있다. 이 과정에서는 이전 진행과정에서의 분류근거를 데이터화 하여 데이터마이닝 분류의 교사학습으로 삼는다. 수치적 근거사항과 항목간 거리차가 입력이 되는 것이다. step1~ step2에서 판단된 사례들이 데이터마이닝 분류수행을 위한 학습패턴으로 동작하고, 이 입력을 이용하여 데이터마이닝을 수행하면 후보항목의 패턴을 조건항으로 갖는 규칙이 생성된다. 후보항목리스트 L_i의 수치와 항목배열별 패턴이 특이한 사례일수록 효과적일 것이다. 이 규칙에 의해 전체 문서 D의 후보리스트 L에 대해 패턴분류가 가능해지게 된다.

3.2.4 step 4 : 유효성 검증에 의한 해석 및 피드백(feedback)

대부분의 분류 실험 사례에서 나타나듯이 분류기의 판별력이 주로 문서개수만으로 평가되고 있다. 그러므로 실제 학습력 부족에서 기인하는 오류 발견이 쉽지 않은데, 많은 양의 학습문서가 반드시 좋은 결과를 보장한다고 볼 수 없기 때문이다. 반면, 데이터마이닝 작업에서는 시스템의 오류와 성능을 평가하기 위해 여러 가지 측정변수와 계산방법이 사용되고 있고, 이때 이전의 분석과정보다 더 큰 분석력과 시간비용이 소요된다. 따라서 분석결과를 해석하여 전단계로 회귀반영할 수 있는 자동화된 시스템의 요구와 전문가의 역할이 대두되고 있다.

Step 4의 과정은 문서 D_i가 분류되기까지 데이터마이닝 프로세스로 진행된 과정을 종합적으로 분석하는 단계이며, 결함허용(fault tolerant)시스템의 접근방식을 응용하여 RTPost 시스템을 안정되게 유지한다. 결함허용시스템은 하드웨어나 소프트웨어 등이 오동작하거나 정보의 오염이 일어나 오류가 발생해도 규정된 기능을 올바르게 수행할 수 있는 시스템을 말한다[20, 21]. 오류가 발생하면 자동복구기능이 작동되거나 다른 기능에 영향을 주지 않도록 설계된 시스템이다. 비행장치와 같이 안정성을 추구하는 복잡한 기계회로에 필수적이다. 결함의 원인으로는 설계상의 실수나, 외부환경으로부터의 영향, 구현의 실수 등으로 분류된다. 결함허용시스템은 결함을 감지(fault detection)하고 발생장소를 검출(fault location)하여 그에 따른 파급효과를 방지(fault isolation) 및 복구(fault recovery)함으로써 시스템을 재구성하는 것을 기본기능으로 한다.

본 연구의 피드백 과정은 RTPost 시스템의 각 진행상태의 출력값으로부터 오류패턴들을 감지하여 분류기와 환경변수들의 결함이나 특성으로부터 안정성을 보장하는 것을 목표로 하고 있다.

이를 위하여 문서를 입력으로 한 1차 분류결과와 후보항목리스트를 입력으로 한 2차 분류결과로 <표 2>와 같이 정의하였다. 각 분류결과의 값과 실제값이 같으면 '1', 틀리면 '0', 미지정문서는 'X'로 표기하였다. 여기에 분류항목을 자질값으로 갖는 2차 결과를 추가한 이유는 그 자체가 분류의 기준이 될 수 있을 만큼 구분력을 지니기 때문이다. <표 2>의 각 단계별 결과값들의 변이(variance)를 통해 어느 단계에서 오류가 있는지 추측할 수 있는데, step 1은 학습문서집합구성과 파라미터 수치, step 2는 거리함수에 의한 기준이 반영되었다고 볼 수 있기 때문이다. step 1과 step 2로부터의 학습패턴들은 step 3에 반영되고 있다.

<표 2>는 출력값들이 기대값으로 진행되는지 검사하고 특수한 오류패턴이 발생하는지 파악하기 위한 상황표이다. 이 값들을 기준으로 분류프로세스의 실행기준과 적합성을

평가하기 위한 척도로 유효성(effectiveness) 함수를 정의하기로 한다. 여기서는 RTPost 시스템의 학습 및 동작상태를 평가하기 위해 문서의 분류결과가 기대값으로 진행되는 패턴과 예외적으로 발생하는 오류패턴을 중심으로 하여 'Good', 'Fair', 'Poor'의 세 가지로 나누어 다음과 같이 정의하였다.

- **정의 8 : Good(di)** 입력문서 d_i 가 경계영역에 있고 다음의 기대값으로 옳게 분류되면 RTPost 시스템이 매우 적절히 동작하고 있다고 간주하고, Good(d_i)에 점수(benefit)를 준다.
- **정의 9 : Fair(di)** 입력문서 d_i가 경계선에 인접하지 않고 옳게 분류되면 RTPost 시스템이 오류가 없이 동작하고 있다고 간주하고, Fair(d_i)에 점수를 준다.
- **정의 10 : Poor(di)** 입력문서 d_i가 경계선이 인접하지 않고 옳게 분류되지 않으면 RTPost 시스템의 학습과 수행과정에 오류가 있다고 간주하고, Poor(d_i)에 벌점(penalty)을 준다.

<표 2>의 각 문서들의 예측된 분류값과 실제값의 변이를 이용하여 위 세 가지 상태에 해당하는 조건을 (그림 7)과 같이 정하였다.

A)~D)의 조건들은 실제값과 예측된 분류값들의 변이로부터 각 세 가지 상태에 가장 영향력 있는 변수를 찾아 조합을 만들고, 이를 다시 단순화시키는 방법을 취하여 정하였다. 우선적으로 입력문서 d_i 의 결과값들 중에서 step 1의 결과값이 실제값과 같은 경우를 'fair', step 1과 step 2의 모든 값이 실제값과 다를 경우를 가장 'poor'한 상태로 판단하였다. 반면, step 1에서는 'X' 즉, 미지정되었으나 이후의 결과값이 실제값과 같으면 'good' 으로 간주하였다. 이는 RTPost 시스템의 프로세스가 의도하는 대로 가장 잘 동작하고 있는 예이다. (그림 7)에서 Good(d_i) =1 은 입력문서 d_i가 조건A와C 를 만족하면 점수1을 준다는 의미이다. 위의 세 가지 상태를 기준으로 RTPost 시스템의 유효도를 측정할 수 있는 함수를 식(5)와 같이 정의하였다.

A : (C_n^1.st3
ep1 = X)
B : (C_n^1.step1 = = True)
C: (C_{n-1}^1.step1 = = True) || (C_n^1.step2 = = True)
D : (C_n^1.step1 = = False) || (C_n^1.step2 ≠ C_n^2)

$$Good(d_i)=\begin{cases}1 & \text{If (A and C)}\\ 0 & \text{otherwise}\end{cases} \quad Fair(d_i)=\begin{cases}1 & \text{If (B)}\\ 0 & \text{otherwise}\end{cases} \quad Poor(d_i)=\begin{cases}1 & \text{If (D)}\\ 0 & \text{otherwise}\end{cases}$$

(그림 7) 유효성평가를 위한 기준 정의

$$E(RTPost)=\frac{1}{N}\left[\sum Good(d_i)\times benefit+\frac{1}{N}\sum Fair(d_i)-\frac{1}{N}\sum Poor(d_i)\times penalty\right] \quad (5)$$

$$benefit=log\,(n)+1.0 \quad (6)$$

$$penalty=log\,(n)+1.5 \quad (7)$$

식 (6)과 식(7)은 'good'과 'poor' 인 상태에 부여된 가중치로서 n은 각 상태에 해당되는 문서빈도수이다. 이때, 'good'과 'poor'한 문서가 같은 빈도로 나타났더라도 'poor'한 쪽에 더 큰 값의 벌점(penalty)을 주었다. 이 값을 이용하여 분류 프로세스가 제대로 동작하는지를 판단한다. 적정값 이하의 유효값을 갖을 경우는 학습과정이 부진했거나 수치분석상 오류가 있다고 판단할 수 있다. 해당범주의 구분력을 높이기 위해 학습문서를 더 많이 보강하거나 수치분석에 사용한 입력변수와 조절상수들을 변경해야 한다. 또한 적정값이 하한값에 가깝도록 낮은 경우 프로세스 전반에 걸쳐 오류가 있다고 판단할 수 있다. 따라서 분류의 가장 초기작업인 목표항목과 경계항목을 정하는 작업부터 다시 살펴 볼 필요가 있다. 예를 들어 검증용 문서개수가 1000일 때 유효값 E(RTPost) 값의 범위는 $-4.5 < E < 4$ 가 된다. 이 때, 'good'인 경우가 없다고 가정하고 'poor'인 경우가 30% 이상이면 E(RTPost)값은 0 미만이 된다.

4. 실험 및 평가

본 실험에서는 오분류가능성이 높은 문서집단에서의 판단력과 정확도를 측정하기 위해서 불확실성이 강하고 직관적인 판단이 어려운 문서집단을 이용하였다. 또한, 분류기의 안정성을 테스트하기 위해 오류문서를 추가한 학습집합에서 비교실험을 수행하였다. 본 논문에서는 생명공학분야의 문헌정보를 대상으로 한 실험을 요약해 보이기로 한다.

4.1실험데이터 및 실험 환경

4.1.1 실험데이터

본 연구의 실험을 위해 망막아세포종(retinoblastoma, RB)에 관한 문서들로서, 'RB' 또는 'pRb' 등의 관련단어를 포함하는 PubMed abstract 문서를 수집하였다. 이 문서에서 Rb로 표현되는 단어는 암(diseace,cancer), 세포주(cell line), 단백질(protein), 유전자(gene), mRNA, 루비듐(ion)등의 다의성을 지니고 있다. retinoblastoma 그 자체로는 암의 한 종류이며 중의적으로 암 환자 조직으로부터 만들어진 세포주(cell line)를 의미하기 한다. 흔히 Rb라고 줄여 말할 때에는 '암' 화되는데 관여하는 유전자, 혹은 단백질을 뜻하기도 한다. 이때의 표현방식은 pRb가 될 수도 있고 Rb가 될 수도 있다. 따라서 이 문서가 말하는 대상이 무엇인지 의미를 파

〈표 3〉 실험문서가 갖는 다의적인 특성

(1) P130 i mediates TGF-beta-induced cell-cycle arrest inn **Rb** mutant HT-3 cells. *(gene)*
(2) The INK4alpha/ARF locus encodes p14(ARF) and p16(INK4alpha), that function to arrest the cell cycle through the p53 and **RB** pathways, respectively. *(protein)*
(3) Many tumor types are associated with genetic changes in the **retinoblastoma** pathway, leading to hyperactivation of cyclin-dependent kinases and incorrect progression through the cell cycle. *(disease,cancer)*
(4) The Y79 and WERI-Rb1 **retinoblastoma**cells, as well as MCF7 breast cancer epithelial cells, all of which express T-channel current and mRNA for T-channel subunits, is inhibited by pimozide and mibefradil with IC(50)= 8 and 5 microM for pimozide and mibefradil, respectively). *(cell line)*

악하기 위해서는 앞,뒤 문장이나 관련 단어들을 각각 살펴보아야 하며 이때, RB gene, RB protein처럼 밝히지 않은 경우에는 이해하는 데 많은 시간비용이 들게 된다.

이는 전체적인 사전정보나 온톨로지정보가 부족한 바이오인포매틱스 분야에서의 개체 명시(entity annotation) 작업과 연결되어 해결해야 할 가장 기본적이며 커다란 문제이며, 자질을 중심으로 한 기존의 분류방법과의 비교 실험에 적당하다.

4.1.2 실험 방법 및 환경

RB 관련문서중 다수는 단백질(P)과 유전자(G), 암(D)에 관한 내용들이고, 이온(I)에 관한 문서는 소수로서 위 항목과 가장 거리가 멀다. 따라서 최우선 목표항목 C={I, others}로 정의하여 분류한 후에 다시 C={P, G, D} 로1대 1(one-against-one)의 분류를 수행하는 것이 효율적이다. 본 실험에서는 이 세 개의 범주에 대해서, 경계항목을 제외한 목표항목을 두 개씩 균등분할하여 S={P1,P2, X1,G1,G2, X2, D1,D2}를 정의하였다. 이 때, 각각의 목표항목마다 60개씩 총300개의 학습문서를 사용하였고, 결과값을 확인한 200개의 테스트문서를 포함하여 총 2000개의 문서로 실험하였다. 실제값이 확인된 200개의 테스트문서는 중의적인 자질의 빈도가 많아서 복잡도가 높거나 내용이 모호한 문서들을 중심으로 구성하였고, 이들의 후보리스트는 유효성 평가시 검증데이터(validation data)로 이용하였다. 그리고 학습문서표본에서 발생할 수 있는 오류로부터의 안정성을 실험하기 위해 학습문서로 부족하거나 부적당한 문서 약10%를 추가하여 이전 결과와 비교하였다.

RTPost 시스템에 의한 마이닝 분류프로세스는 반복적인 분석작업과 학습문서구성을 자동화하여 효과적으로 수행할 수 있도록 구성하여 리눅스환경에서 JAVA로 구현하였다. 이 때, 1차 문서분류를 위한 알고리즘은 Naïve Bayesian과 SVM을 적용하였고, 2차 패턴분류를 위한 알고리즘은 블랙박스의 형태로 설명력은 부족하지만 성능면에서는 우수한 신경망(Neural Network)을 적용하였다. 1차 분류시1순위 항목으로 지정하기 위한 수치조건은 min_support=100(bytes), min_value=0.6, diff_value=0.2로 정하였다. 그리고 유효도의 적정값은 0.5로 하고 한번의 피드백을 수행하여 결과를 분석하였다.

〈표 4〉 목표항목 및 학습문서집합의 구성

목표 항목 및 세부항목 구성			학습문서 개수		
목표항목 (C)	세부 항목(S)	경계 항목(X)	학습 문서	오류포함 (약 10%)	총 (300, 318)
P(protein)	P1		30	5	60(36)
	P2		30	1	
	ㄷ	X1	60	0	60
G(gene)	G1		30	3	60(36)
	G2		30	3	
		X2	60	0	60
D(disease,canse)	D1		30	6	60(36)
	D2		30	0	

4.2 평가방법

실제범주와 추정분류의4가지 값(TP,TN,FP,FN)으로 구성된 정오분류표(confusion matrix)를 이용하여 평가한다. 여기서 'positive | nagative'는 '양성|음성'으로 추정됨을 나타내고, 'true | false ' 는 추정결과가 '참 | 거짓'이냐를 나타낸다. 따라서 정확도(accuracy)는 (TP+TN)/total, 오류율(misclassificationn error)은 (FP+FN)/total로 산출된다.

한편, 2*2 matirx에서는 정확율(precision)과 재현율(recall) 이외에 정확도를 평가하는 기준으로 민감도(sensitivity) 및 특이도(specificity)를 함께 분석하는 경우가 종종 있다. 이 때 민감도의 의미는 실제 양성인 경우를 양성으로 예측해낸 정확도를 나타내며, 특이도는 그 반대로 실제 음성인 경우를 음성으로 옳게 예측해 냈는지를 의미한다. 즉, 양성|음성의 예측값과 실제값이 일치하는 정도를 평가하며, 각 산출식은 <표 5>와 같다.

〈표 5〉 정오분류표(confusion matrix)와 측정수치

		Actual Class	
		1	0
Predicted Class	1	TP	FP
	0	FN	TN

민감도(sensitivity) = $\frac{TP}{TP+FN}$, 특이도(specificity) = $\frac{TN}{TN+FP}$

정확율(precision) = $\frac{TP}{TP+FP}$, 재현율(recall) = $\frac{TP}{TP+FN}$ (=positive predict power)

이 값을 별도로 산출하여 분석하는 이유는 같은 정확도를 갖는다 하더라도 민감도와 특이도가 다를 수 있으며, 경우에 따라서는 높은 민감도나 높은 특이도가 관건이 될 수 있기 때문이다. 본 실험의 평가에서는 실제값이 확인 된200개의 문서를 대상으로 정확율과 오류율을 확인하였다.

4.3 실험결과 및 평가

4.3.1 실험결과

본 실험에서는 세 개의 범주에 대한 분류결과로서 3*3 matrix의 분류표를 얻을 수 있다. 아래 표에서는 예측결과와 실제값을 확인하여 각 범주의 예측력(positive predict power)을 계산하였다. <표 6>과 <표 7>은 각각 올바른 학습문서와 10%의 오류문서를 추가하여 실험한 결과를 보이고 있다. 오류문서가 없는 <표 6>에서도 기존방법과 제안방법의 차이를 볼 수 있다. 분류력이 높다고 평가받는 SVM의 경우에도 좋지 않은 결과를 보인다. 이는 실험문서가 가진 문제점, 즉, 여러 자질특성을 공유하고 있으므로 직관적으로 분류하기 어렵다는 특성에서 기인하는 결과로 추정할 수 있다.

반면, 제안방법에 의한 결과는 상대적으로 높은 정확도를 나타냈다. 특히, 'gene, dna, mRNA'를 주요자질로 하는 'gene'범주와 'cancer'를 주요 자질로 한 'disease' 범주의 예측력이 높았고, 상대적으로 'protein' 범주의 예측율이 낮은 편인데, 이는 'protein' 범주의 문서들의 가진 자질비율과 복잡도가 높음을 반영한다고 볼 수 있다.

<표 7>은 오류문서 10%를 추가하여 분류기의 안정성을 실험한 결과이다. 예상한 대로 기존의 방법에서는 정확도가 현저히 감소되었음 알 수 있다. 이는 특이하고 주요한 자질들이 뒤섞인 오류문서로 인하여 자질에 의한 범주간 구분능력이 상쇄되었음을 의미한다. 실제적으로 SVM은 학습표본에 민감하여 오류표본 증가율에 따라 성능이 급속히 저하되

는 것으로 알려져 있는데, 이 실험을 통해서도 확인할 수 있었다. Naïve Baysian 역시 학습표본에 큰 영향을 받지 않는 상당히 안정적인 알고리즘이긴 하지만, 'disease' 범주의 예측율이 74%에서 17%로 떨어진 점으로 볼 때 확률적인 계산과 단순지정방식에 의한 정확도의 한계를 보여주고 있다.

반면, 제안방법은 이전과 큰 차이를 보이지 않으면서 학습표본상의 오류로부터 높은 안정성과 상대적으로 낮은 오류율을 보이고 있다. 이는 경계항목을 기준으로 한 후처리 분석과, 유효성평가에 의한 재학습으로 오류문서에 의한 영향을 완화시켜주었음으로 분석된다. 공통적으로 기존방법과 제안방법의 각 범주의 예측력이 감소했지만 특히 'disease' 범주의 예측력 감소가 가장 두드러졌음을 볼 수 있다. 이는 'disease'의 대표자질들이 단 몇 개의 오류문서에 의해서 그 효력이 상쇄되었음을 의미한다. 이 결과는 주요자질들에 의존하는 정보력과 구분능력은 학습표본에 의존하고 있으며, 이는 분류결과의 신뢰도에 매우 큰 영향을 미치고 있음을 강조하고 있다. (그림 8)에서 알 수 있듯이 제안방법은 다양한 분류 알고리즘과 분류기의 특성, 적용환경에 영향을 덜 받는 안정된 결과를 보이고 있다. 따라서 정확도 향상과 더불어 자동화 방식에서 예상할 수 있는 오류에 민감하지 않다는 부수적인 효과를 기대할 수 있다는 점으로 해석된다.

〈표 6〉 기존 방법과 post processing후의 성능 분석결과

performance / method	Accuracy	Protein Predict P.	Gene Predict P.	Disease Predict P.	Misclassification rate
Naïve Baysian	0.69	51%	82%	74%	31%
SVM	0.74	64%	83%	76%	29%
RTPost Algorithm (Naïve Baysian)	0.89	81%	94%	92%	11%
RTPost Algorithm(SVM)	0.91	88%	91%	94%.	8%

〈표 7〉 오류가 포함된 학습문서집합에 의한 분석 결과 (오류 문서 10%)

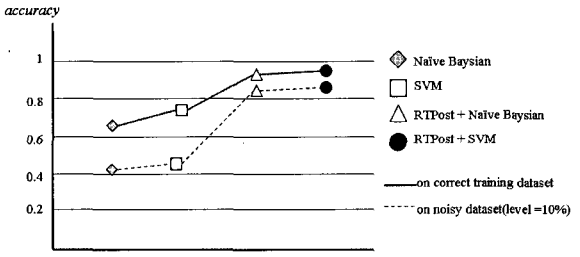
performance / method	Accuracy	Protein Predict P.	Gene Predict P.	Disease Predict P.	Misclassification rate
Naïve Baysian	0.45	52%	65%	17%	55%.
SVM	0.47	54%	61%	26%	64%
RTPost Algorithm (Naïve Baysian)	0.85	84%	92%	75%	15%
RTPost Algorithm(SVM)	0.87	87%	91%	81%	11%

4.3.2 평가 및 토의

위의 실험에서는 <표 4>에서와 같이 목표항목을 2개의 세부항목으로 균등분할하였다. 그러나 범주가 다루는 내용의 다양성과 복잡도에 비례하여 분할하는 방법도 고려해 볼 수 있다. 일반적인 분류에서는 문서에 대응하는 분류항목이 1:1을 추구하고 있고 이번 실험에서도1:1로 하였으나, 해당 범주와 문서의 사이즈가 크고 복합적이고 다양한 내용을 내포하고 있을 경우에는 여러 분류항목으로 할당되게 하는 것도 고려해 볼 수 있다.

현실세계에서 자동문서분류기법을 사용할 때, 해당 영역의 학습데이터에 적지 않은 오류 문서가 포함되거나 혹은 최하위 범주까지 분류하는 작업은 일반적인 일이다. 특히 위의 결과에서 보는 바와 같이 학습표본상에서 오류의 포함여부는 성능에 많은 영향을 끼친다. 그리고 이 실험에 앞서 세부분류를 위한 판단력 테스트로 뉴스그룹 데이터를 사전실험에 이용하였는데 'comp' 그룹과 'sci'그룹등의 대분류 구분력에서는 기존방법과 차이가 없었으나 그 하위 그룹들에서는 차이를 보였고, 'comp.sys.mac.hardward'와 'comp.sys.ibm.hardware' 같은 최하위 그룹들에서는 현저한 차이를 볼 수 있었다.

일반적으로 기계학습분야에서는 분석하려는 데이터가 무엇이고 어떠한 자질특성을 이용했느냐가 일반적인 성공요인이다. SVM에서는 어떤 커널함수를 적용했고, 변수값들은 어떻게 주어졌으며, 다원분류를 해결하기 위해 어떠한 방법을 사용했는지가 결과에 가장 영향을 미치는 요소이다. RTPost 시스템의 성능에 영향을 줄 수 있는 요인으로는 경계항목의 구성과 범주할당 계산에 쓰인 파라미터들의 상한값과 하한값, 그리고 전체 프로세스의 판단력을 평가하는 유효성 계산 함수를 들 수 있다. 이 요소들은 이전의 결과를 보증(guarantee)하고 보완(repair)하는 의미로 사용된었다.



(그림 8) 오류문서 포함여부에 따른 RTPost 시스템의 성능비교

5. 결론 및 향후연구

문서의 내용이 광범위한 주제를 다루고 있거나 동일한 수준의 주제를 다루지만 더 많은 단어들을 사용하면 그 복잡도가 높아진다. 문서의 복잡도가 높은 문서일수록 오분류율이 높기 마련이며, 이런 문서들의 자동분류 성능향상을 위해 분류알고리즘의 최적화 연구를 주로 하여 학습문서의 자질표현방법을 개선한 연구들이 있어왔다.

본 연구에서는 실세계에서 발생하는 오류요인에서도 높은 정확성을 유지할 수 있는 자동문서분류를 위해 학습문서집합 구성방법, 최종항목 할당방법을 개선하여 데이터마이닝 프로세스에 따른 분류모형을 제시하고자 하였다. 그리고 기존 방법과 비교하여 그 성능이 어떻게 달라지는가를 평가하는 실험을 수행하였다. 이는 중의적 특성을 갖는 복잡한 문서가 각 범주간의 경계상에 놓일 수 있다는 점과 현실세계에서 자동문서분류를 수행할 때 대부분의 오류가 소극적인 지정방식과 오류결과에 대한 대응부족에서 기인하고 있다는 문제에서 출발하였다. 전자는 사전정보(back-of-word)를 위주로 한 예측능력의 한계를 경계항목을 포함한 후보항목들의 분석을 통해 개선할 수 있었고, 후자는 유효성 검사에 의한 피드백 지침을 통해 개선할 수 있었다. 이는 재학습과

전문가의 개입여부를 정하는 데도 도움이 된다. 또한 문서 표현(document representation)방법으로서 사전정보 뿐만 아니라, 분류가능성 있는 항목의 리스트 정보를 이용하였다. 항목의 리스트 정보는 문서자체를 분류 가능한 항목들값을 기준으로 데이터화 시킨 것이므로, 분류 문제에 있어서 문서를 표현하는 또 다른 방법이 될 수 있다. 이 리스트 정보를 이용하여 분류 결과를 얻어내어 RTPost 시스템 평가와 피드백과정에만 이용하였는데, 실험한 결과 실제 분류 문제에 적용시키는 것도 무리가 없어 보였다.

본 실험에서 이용한 유효성 함수는 'good, fair, poor'의 세 가지 상태에 따라 간단히 정의하였으나, 전체 프로세스의 비용을 높이지 않는 선에서 데이터마이닝에서 이용하는 복합적인 평가지표들을 응용하는 방법을 고려해 볼 수 있다.

마지막으로 본 연구에서 이용한 두 문서표현방식의 장점을 취하여 보다 영향력 있는 좋은 패턴을 선택하여 실험하는 것과 분류프로세스의 판별력을 검사할 수 있는 간결한 유효성함수 설계에 대한 내용을 향후 과제로 남겨둔다.

참 고 문 헌

[1] R. Agrawal, R. Bayardo, and R. Srikant, "Athena: Mining-based Interactive Management of Text Databases," In Proceedings of the 7th International Conference on Extending Database Technology, pp.365-379, 2000.

[2] Yiming Yang. "An Evaluation of Statistical Approaches to Text Categorization," Journal of Information Retrieval, Vol.1, No.1, pp.67-88, 1999.

[3] Zijian Zheng. "Naïve Bayesian Classifier Committees," In Proceedings of European Conference on Machine Learning, pp.196-207, 1998.

[4] Yiming Yang and J. O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization," In Proceedings of the 14th International Conference on Machine Learning, pp.42-420, 1997.

[5] David D. Lewis and Jason Catlett. "Heterogeneous Uncertainty Sampling for Supervised Learning," In Proceedings of the 11th international Conference on Machine Learning, pp.148-156, 1994.

[6] Pedro Domingos and Michael Pazzani. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", In Proceedings of the 13th International Conference on Machine Learning, pp.105-112, 1996.

[7] Kim S.B., H.C.,Rim, "Recomputation of Class Relevence Score for Improving Text Classification," In Proceedings of Conference of CICLing(Computational Linguistics and Intelligent Text Processing), Lecture Note in Computer Science, Vol.2945, pp.580-583, Feb., 2004.

[8] Ko, Y.J., J.Y., Seo, "Using the Feature Projection Technique based on a Normalized Voting Method for Text Classification," Information Processing & Management, Pergamon-Elsevier Science, Vol.40, No.2, pp.191-208, Mar., 2004.

[9] 김제준,김한구, "베이지언 문서분류시스템을 위한 능동적 학습 기반의 학습문서집합 구성방법", 한국정보과학회 논문지, Vol.29, No.12, 2002. 12.

[10] Wilson, D.,R., et al "Reduction Techniques for Exemplar-based Learning algorithms," Machine Learning, Vol.38, No.3, pp.257-286, 2002.

[11] T.Joachims, "Text categorization with support vector machines: learning with many relevant features," In Proceedings of ECML-98, 10th European Conference on Machine Learning, pp.137-142, 1998.

[12] C., Cortes and V., Vapnik, "Supprot Vector Network", Machine Learning, Vol.20, pp.273-297, 1995.

[13] D. Koller and S. Tong. "Active learning for parameter estimation in Bayesian networks," In Neural Information Processing Systems, 2001.

[14] M. Hasenager. "Active Data Selection in Supervised and Unsupervised Learning," PhD thesis, Technische Fakultat der Universitat Bielefeld, 2000.

[15] Dagan, I. And A.Itai, "Word Sense Disambiguation using a second language monolingual corpus," Computational Linguistics, 20(4), December, 1994.

[16] Hatzivassiloglou, V., P.A. Duboue, and A.Rzhetsky. "Disambiguating Proteins, Genes and RNA in Text: a Machine Learning Approach". Bioinformatics Vol.17, pp.S97-106, 2001.

[17] Tateishi, Y., T. Ohta, J. Tsujii, "Building an Annotated Corpus in the Molecular-Biology Domain," In Proceedings of COLING 2000 Workshop on Semantic Annotation and Intelligent Content, pp.28-34, 2000.

[18] S. B. Cho, "Ensemble of structure adaptive self-organizing maps for high performance classification," Inforrmation Science, Vol.123, No.1-2, pp.103-114, 2000.

[19] W.N. Street, and Y. S. Kim, "Streaming ensemble algorithm(SEA) for large-scale classification," Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.377-382, San Francisco, California, 2001.

[20] B. Krishnamachari and S. Iyengar, "Distributed Bayesian Algorithms for Fault-Tolerant Event Region Detection in Wireless Sensor Networks," IEEE Transactions on Computers, Vol.53, No.3, pp.241-250, March, 2004.

[21] D. K. Pradhan, ed., Fault-Tolerant Computer System Design. Prentice Hall Inc., 1996.

최 윤 정

e-mail : cris@ewhain.net

1997년 서원대학교 전자계산학과(학사)

2001년 이화여자대학교 컴퓨터학과(석사)

2001년~현재 이화여자대학교 컴퓨터학과 박사과정

관심분야 : 인공지능, 텍스트마이닝, 데이터마이닝, 바이오인포매틱스

박 승 수

e-mail : sspark@ewha.ac.kr

1974년 서울대학교 수학과(학사)

1976년 한국과학기술원 전산학(석사)

1988년 미국 텍사스대학 전산학(박사)

1988년~1991년 미국 켄사스대학 컴퓨터학과 조교수

1991년~현재 이화여자대학교 컴퓨터학과 교수

관심분야 : 인공지능, 데이터마이닝, 바이오인포매틱스