

# A Novel Posterior Probability Estimation Method for Multi-label Naive Bayes Classification

Hae-Cheon Kim\*, Jaesung Lee\*\*

## Abstract

A multi-label classification is to find multiple labels associated with the input pattern. Multi-label classification can be achieved by extending conventional single-label classification. Common extension techniques are known as Binary relevance, Label powerset, and Classifier chains. However, most of the extended multi-label naive bayes classifier has not been able to accurately estimate posterior probabilities because it does not reflect the label dependency. And the remaining extended multi-label naive bayes classifier has a problem that it is unstable to estimate posterior probability according to the label selection order. To estimate posterior probability well, we propose a new posterior probability estimation method that reflects the probability between all labels and labels efficiently. The proposed method reflects the correlation between labels. And we have confirmed through experiments that the extended multi-label naive bayes classifier using the proposed method has higher accuracy than the existing multi-label naive bayes classifiers.

▶ Keyword: Multi-label Classification, Naive Bayes Classifier, Posterior Probability Estimation, Label Dependency

## 1. Introduction

다중 레이블 분류는 데이터를 입력받았을 때, 그와 관련된 여러 레이블이 무엇인지 결정하는 문제이다. 대표적으로 글에 포함된 단어를 분석하여 특정 문장이 표현하는 감성을 인식하는 텍스트 감성 인식 문제 [14, 17, 24], 이미지나 동영상에 포함하는 정보를 분석하여 정보가 가진 특징을 유추하는 이미지 태그 문제 [15], 효소가 포함하는 요소의 종류와 개수 등을 분석하여 효소의 생물학적 활동을 추측하는 유전자 기능 추측 문제 [11] 등의 연구들에 다중 레이블 분류가 사용되었으며, 관련된 분류기를 활용하여 예술 작품 혹은 인공지능 인문학으로 접목하는 연구 또한 시도되고 있다 [25, 26, 27].

다중 레이블 분류 수행을 위해 널리 활용되는 방법은 단일 레이블 분류 기법을 확장하는 것이다. k-근접 이웃 분류기

(k-Nearest neighbor, k-NN)를 확장한 ML-kNN [3], 나이브 베이즈 분류기 (Naive bayesian, NB)를 확장한 MLNB [4], 그리고 서포트 벡터 분류기 (Support vector machine, SVM)를 확장한 ML-SVM [15, 17]이 그 예이다. 이 중 나이브 베이즈 분류 기법은 베이즈 이론을 기반으로 사후확률을 추정하고, 사후확률이 가장 높은 레이블을 결과로 출력하는 기법으로, 추측 속도가 빠르며 높은 정확도를 보이는 분류기법이다.

한편, 단일 레이블 분류 기법을 다중 레이블 분류 기법으로 확장할 때 크게 3가지 방식을 적용한다. 첫 번째는 레이블마다 분류기를 할당하여 각자 레이블이 나타날 확률을 계산하는 이진 연관성(Binary relevance) 방법이다. 두 번째는 레이블 집합을 단일 레이블로 치환하여 분류기에 적용하는 레이블 멱집

• First Author: Hae-Cheon Kim, Corresponding Author: Jaesung Lee

\*Hae-Cheon Kim (code.bug.station@gmail.com), School of Computer Science and Engineering, Chang-Ang University

\*\*Jaesung Lee (curseor@cau.ac.kr), School of Computer Science and Engineering, Chung-Ang University

• Received: 2018. 03. 20, Revised: 2018. 04. 23, Accepted: 2018. 05. 28.

• This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A6A3A01078538).

• This paper is expanded from proceeding(HC Kim and J Lee, 2018, "Improving Accuracy of Multi-label Naive Bayes Classifier") in the 57th winter conference 2018 of the Korea Society of Computer and Information.

합(Label powerset) 방법이며, 세 번째는 지금까지 추측한 레이블 집합 결과를 다른 레이블 예측 시에 연속적으로 반영하여 결과를 추측하는 분류기 체인(Classifier chains) 방법이다.

각 방법을 통해 단일 레이블 분류기들은 다중 레이블 분류기로 확장할 수 있지만, 기존 연구에서는 사후확률 추정 시 레이블 간의 연관성을 고려하지 않았다 [1, 4, 14]. 이에 본 논문에서는 위의 문제를 해결하기 위해 다중 레이블 나이브 베이즈 분류기의 새로운 사후확률 추정 식을 제안하고자 한다. 이후 제안한 확률 추정 식을 다중 나이브 베이즈 모델에 적용하고자 한다.

## II. Related Work

다중 레이블 분류는  $d$ 차원 데이터  $X=[x_1, x_2, \dots, x_d]$ 을 입력받았을 때,  $q$ 개의 레이블 중 관련된 레이블들을 포함하는 분류 집합  $L=\{l_i | 1 \leq i \leq q\}$ 으로 사상하는 문제이다. 레이블 집합  $L$ 은  $i$ 번째 레이블을 포함하면  $i$ 번째 숫자를 1, 그 외에는 0으로 하여 이진 벡터  $V \in 2^q$ 으로 표기한다. 예를 들어, 총 레이블 개수가 5개이고  $L=\{l_1, l_3, l_4\}$ 이면,  $L$ 은  $[1, 0, 1, 1, 0]$ 이다.

### 1. Naive Bayes Classifier

나이브 베이즈 분류기는 학습 데이터를 이용하여 사전확률을 계산하고, 새로운 입력을 받았을 때 베이즈 정리를 기반으로 최대 사후확률로 계산되는 레이블을 추측하는 것이다 [4]. 식 (1)은 베이즈 정리를 통해  $d$ 차원 입력  $x$ 가 주어졌을 때 추측 레이블이  $l$ 일 확률을 추정하는 식이다.

$$P(Y=l|X=x) = \frac{P(X=[x_1, \dots, x_d] | Y=l) P(Y=l)}{P(X=[x_1, \dots, x_d])} \quad (1)$$

$x$ 의 모든 경우가 학습 데이터 안에 존재하면 식 (1)을 완벽하게 계산할 수 있다. 하지만 데이터의 수집에 한계가 있기 때문에 입력 차원의 데이터는 레이블의 선택과 조건부 독립이라는 전제를 적용하여 확률 추정 식을 다음과 같이 느슨하게 (Naive) 변형한 식으로 사후확률을 구한다. 식 (2)는 나이브 베이즈 분류기의 사후확률 계산식이다. 이때, 데이터의 확률모델로 표준분포, 혹은 다변주 분류분포임을 가정한다.

$$\begin{aligned} & \arg \max_l P(Y=l|X=x) \\ & \propto \arg \max_l \prod_{i=1}^d P(X_i=x_i | Y=l) P(Y=l) \dots \quad (2) \end{aligned}$$

위 사후확률 추정 식 (2)을 이용하여 최대 사후확률을 도출하는 레이블 분류기를 나이브 베이즈 분류기라 한다.

## 2. Extension to Multi-label Classification

다중 레이블 분류는 단일 레이블 분류와 다르게 여러 레이블이 포함된 집합을 결과로 출력해야 한다. 그러므로 식 (3)을 만족하는 벡터  $V$ 를 찾는 것을 목표로 한다.

$$\arg \max_{V \in \{0, 1\}^q} P(Y=V|X=[x_1, x_2, \dots, x_d]) \dots \quad (3)$$

다중 레이블 분류를 수행하기 위해 대표적으로 이진 연관성, 레이블 먹집합, 그리고 분류기 체인 방법을 활용한다.

### 2.1. Binary Relevance

이진 연관성 방법은 각 레이블의 결과 포함 여부를 각각 할당된 분류기를 통해 판단하고, 결과를 취합하여 최종 집합을 추측하는 과정을 거친다. 나이브 베이즈 분류기에 이진 연관성 방법을 적용할 경우, 레이블마다 식 (4)을 이용하여 사후확률을 추정하여 각 레이블의 포함 여부를 판별하고,  $q$ 개의 레이블 분류기의 결과를 취합하여 레이블 집합을 찾아낸다.

$$\arg \max_{V_q \in \{0, 1\}^q} P(Y_q=V_q | X=[x_1, x_2, \dots, x_d]) \dots \quad (4)$$

이진 연관성 나이브 베이즈 분류기는  $q$ 개의 단일 나이브 베이즈 분류기를 생성하여 병렬로 연결하는 구조로 레이블의 개수에 따른 확장이 매우 쉽다는 장점이 있다. 하지만 각자 레이블의 포함 여부만을 판단하므로 각 레이블 간의 관련성을 계산 상에 고려하지 않는다. 따라서 실제로 존재할 수 없는 레이블 집합을 추측하는 상황이 발생할 수 있어 추측 정확도가 떨어지는 문제점을 가진다.

### 2.2. Label Powerset

레이블 먹집합 방법은 다중 레이블 집합을 단일 레이블 값으로 치환하여 단일 레이블 분류기가 처리할 수 있는 데이터로 바꾸어 레이블 추측에 사용하는 방법이다. 추측이 끝나면, 이후 단일 레이블 값을 레이블 집합으로 환원시켜 다중 레이블 분류를 수행한다. 레이블의 집합이 데이터 상으로 0과 1로 저장됨을 이용해 레이블의 집합을 2진법으로 바꾸어 단일 레이블로 치환하는 방식이 대표적으로 사용된다. 나이브 베이즈 분류기에 레이블 먹집합 방법을 적용하면, 식 (3)을 이용하여 위의 2진법 방식으로 표현한 레이블을 찾아낸 후, 2진법 치환 방식을 역으로 실행하여 레이블 집합을 찾아낸다.

레이블 먹집합 나이브 베이즈 분류기는 단일 나이브 베이즈 분류기와 구조가 같다. 그리고 각 레이블 집합의 경우의 수마다 각자 다른 단일 레이블 값으로 치환되므로, 레이블 간의 연관성을 고려할 수 있는 장점을 가진다. 하지만 데이터 집합 안에서 각 레이블 값의 출현빈도가 매우 떨어져 각 레이블당 학습 데이터의 수가 감소하는 현상이 발생하며, 이로 인해 추측 정확도가 떨어지는 문제점을 가진다.

### 2.3. Classifier Chain

분류기 체인 방법은 다른 레이블들의 포함 여부를 입력 데이터에 포함해 레이블 추측 시 도움을 줄 수 있도록 추측 결과를 연쇄적으로 다음 입력에 반영하는 방법이다. 추측이 끝난 레이블의 포함 여부를 전부 데이터로 포함해 남아있는 다른 레이블을 추측할 때 반영하는 방법 [5, 6]이 대표적이며, 이후 계산 효율성을 높이기 위해 베이즈 네트워크(Bayesian network)로 레이블 연관성을 반영하기도 한다 [1, 2, 17].

나이브 베이즈 분류기에 분류기 체인 방법을 적용할 경우, 각 레이블의 추측 순서에 따라 각 분류기가 연쇄적으로 연결되는 구조를 가진다. 레이블의 포함 여부가 결정된 레이블의 결과는 다음 레이블의 포함 여부를 추측할 때 사용된다. 이때 각 레이블의 포함 여부를 나이브 베이즈의 방식으로 사후확률을 추정하여 이를 비교한다. 식 (5)은 입력에 다른 레이블의 결과를 포함하여 사후확률을 추정하는 식이다.

$$\operatorname{argmax}_{V_q \in \{0,1\}} P(Y_q = V_q | X = [x_1, x_2, \dots, x_d], Y_{[1:q-1]} = V_{[1:q-1]}) \dots (5)$$

이 방법은 특정 레이블의 포함여부를 데이터로 입력받아 추측에 반영하기 때문에 지금까지의 확장 방법보다 레이블 간의 연관성을 잘 반영하여 정확도가 높게 나타나는 장점을 가진다. 이후 순서 효과를 줄이기 위하여 무작위로 순서를 뽑아 여러 번 확률 추정을 시도하고 평균값을 최종 판정 확률로 계산하는 방식인 ECC [18]가 연구되었으며, 연관성을 최대한 계산하여 방향성 비순환 그래프 (Directed acyclic graph) [1] 또는 베이즈 네트워크로 표현하여 [2, 16] 레이블 선택순서를 정하는 방식을 도입하기도 했다. 그러나 레이블 체인의 순서에 따라 분류 결과가 달라질 수 있는 문제점을 가진다.

## III. Proposed Method

기존 다중 레이블 나이브 베이즈 분류기들은 레이블 구조 변경 시 수반되는 문제들로 인해 정확도 상승에 한계를 가지고 있다 [3]. 이에 본 논문에서는 레이블의 구조를 변경하지 않으며 다중 레이블 클래스 분류를 수행하는 새로운 확률 추정식을 제안한다. 다중 나이브 베이즈 분류기는 데이터  $X = [x_1, x_2, \dots, x_d]$ 가 주어졌을 때, 훈련 데이터로부터 사전확률을 계산하고, 이를 기반으로 사후확률이 가장 높을 것으로 추정되는 집합 벡터  $V_L \in \{0,1\}^q$ 을 찾아내는 과정을 거친다. 식 (6)은 그 과정을 수식으로 보인 것이다.

$$\operatorname{argmax}_{V_L \in \{0,1\}^q} P(Y = V_L | X = [x_1, x_2, \dots, x_d]) \dots (6)$$

식 (6)의 확률 식은 조건부 확률식의 정의에 의해 교집합으

로 전개할 수 있다. 이때 분모에 존재하는  $X$ 에 대한 확률 식은 입력으로 받는 값이므로 특정 상수로 고정된다. 이 사실을 이용하여 식 (6)을 식 (7)로 전개한다.

$$\operatorname{argmax}_{V_L \in \{0,1\}^q} P\left(\bigcap_{k=1}^d X_k = x_k, \bigcap_{l=1}^q Y_l = v_l\right) \dots (7)$$

이후, 식 (7)의 교집합 부분 중 레이블 하나를 선정하고, 선택한 레이블이 등장할 확률 식과 선택한 레이블에 대한 조건부 확률 식의 곱으로 치환하여 식 (8)을 전개한다.

$$\operatorname{argmax}_{V_L \in \{0,1\}^q} P(Y_1 = v_1) \cdot P\left(\bigcap_{k=1}^d X_k = x_k, \bigcap_{l=2}^q Y_l = v_l \mid Y_1 = v_1\right) \dots (8)$$

이제 특정 레이블의 존재 여부가 확인됐을 때 남은 레이블들이 각각 포함 여부를 따지는 사건이나 각각의 입력 값이 조건부 독립이라는 가정을 식 (8)에 적용하여 식 (9)을 전개한다.

$$\operatorname{argmax}_{V_L \in \{0,1\}^q} P(Y = V_L | X = [x_1, x_2, \dots, x_d]) \\ \propto P(Y_1 = v_1) \prod_{k=1}^d P(X_k = x_k | Y_1 = v_1) \prod_{l=1}^q P(Y_l = x_l | Y_1 = v_1) \dots (9)$$

식 (9)은 전개 중에 선정한 레이블 1개를 기준으로 사후확률을 추정하는 식이다. 위의 식 (9)을 그대로 사용하는 경우, 어떤 레이블을 선택하느냐에 따라 확률 추정치가 변동하게 된다. 위 문제점을 해결하기 위해, 본 연구에서는 모든 레이블을 한 번씩 선정하여 위 식 (9)에 따라 사후확률을 계산하고, 모든 확률 추정치들을 모아 평균을 내어 최종 추정확률을 도출했다. 식 (10)은 본 논문에서 제안하는 사후확률 추정 수식을 나타낸 것이다.

$$\propto \frac{1}{q} \sum_{l=1}^q \left\{ P(Y_l = v_l) \prod_{k=1}^d P(X_k = x_k | Y_l = v_l) \prod_{i=1}^q P(Y_i = x_i | Y_l = v_l) \right\} \dots (10)$$

식 (10)을 통해 각 레이블을 하나씩 기준으로 세우고, 입력 데이터가 나타날 사후확률을 추정함과 동시에 다른 모든 레이블이 결과에 포함되거나 포함되지 않을 사후확률을 동시에 추정하여 최종적으로 더욱 정확하게 사후확률을 추정할 수 있게 된다. 이후 모든 레이블을 각각 기준으로 세워 계산한 확률들을 모아 평균을 내어, 모든 레이블 간의 상관관계를 사후확률에 반영할 수 있게 된다.

## IV. Experimental Results

본 논문에서 제시한 사후확률 추정 식이 적용된 다중 레이블 나이브 베이즈 분류기의 성능이 이전의 다중 레이블 나이브 베이즈 알고리즘들보다 성능의 향상이 있는지 검증하고자 정확도 비교 실험을 설계했다.

#### 4.1. Experimental Settings

본 실험에서 사용할 데이터들은 Arts, Health [7], Bibtex [8], Cal500 [9], Emotions [10], Enron, Slashdot, Language Log [20], Genbase [21], Medical [22], Scene [15], Tmc2007 [23], Yeast [11]의 총 13가지의 데이터로 모두 다중 레이블 클래스 분류 연구에서 널리 활용되는 데이터들이다. Table 1.은 각 데이터의 특성을 정리한 것이다. Table 1.의 Domain은 본래 데이터의 형태, Pattern은 채집한 건본의 개수, Features는 각 건본이 가지는 특징 개수, Labels는 추측 가능한 레이블 종류의 수를 나타낸다. 본 논문의 제안 분류기는 다변수 다항분포를 사용하므로, 연속적인 수치로 이루어진 데이터는 LAIM 이산화 방법을 거쳤다 [12].

Table 1. Characteristics of employed data sets

Data set	Domain	Patterns	Features	Labels
Arts	Text	7484	1157	26
Bibtex	Text	7395	1836	159
Cal500	Music	502	68	174
Emotions	Emotion	593	72	6
Enron	Text	1702	1001	53
Genbase	Biology	662	1185	27
Health	Text	9205	1530	32
LLog	Text	1460	1004	75
Medical	Text	978	1494	45
Scene	Image	2407	294	6
Slashdot	Text	3782	1079	22
Tmc2007	Text	28596	981	22
Yeast	Biology	2417	103	14

본 실험의 비교 대상은 나이브 베이즈 분류기에서 확장된 다중 레이블 분류기 중에 2가지를 선정했다. 첫 번째는 이진 연관성 나이브 베이즈 분류기 (MLNB-BR)을, 두 번째는 앙상블 기법을 적용한 나이브 베이즈 연쇄 분류기 (MLNB-ECC)을 선정했다.

MLNB-BR 분류기는 각 레이블을 독립적이라고 가정하고 레이블마다 사후확률을 추정하여 레이블마다 결과 포함 여부를 결정한다. 그래서 레이블 간의 연관성을 고려하지 못하는 한계를 지니고 있다. 그래서 MLNB-BR을 첫 번째 비교 알고리즘으로 선정했다. 다른 분류기 MLNB-ECC는 레이블을 무작위로 선택하여 순서를 정하고 그 순서대로 레이블을 추측하되, 지금까지 추측한 레이블들의 결과를 다음 레이블의 포함 여부를 추측할 때 입력 데이터로써 활용하여 사후확률을 추정하는 방식을 지녔다. 이런 과정을 여러 번 반복하여 여러 확률을 얻어내고, 이를 평균으로 내어 최종 사후확률을 계산한다. 이런 방식은 레이블 간의 관계를 고려하지만, 레이블의 선택 순서에 따라 추정된 사후확률이 불안정하다는 문제점을 지니고 있다. 그래서 MLNB-ECC를 두 번째 비교 알고리즘을 선정했다.

이렇게 선정된 2가지 분류기와 본 논문에서 제안한 사후확률 추정 식을 적용한 나이브 베이즈 분류기를 같은 데이터로 훈련하게 시키고 실험하여 각자의 성능을 비교할 것이다. 이 실험에서 사용된

비교 알고리즘들은 각자 제안 논문의 구현을 따랐으며 [4, 18] 앙상블 기법 분류기의 시도 횟수는 10번으로 설정했다.

#### 4.2. Validation and Performance Measurement

검증은 교차검증을 30번 반복하여 이루어진다. 시도마다 건본의 80%를 무작위로 선택하여 학습 데이터로 정하고, 남은 20%를 평가 데이터로 정한다. 시도마다 학습 데이터로 각 알고리즘을 학습하고, 평가 데이터를 입력하여 나온 레이블 집합을 실제 평가 데이터의 레이블 집합과 비교한다.

이때 이번 검증에서 사용하는 성능 측정함수는 총 3가지이다. 이때 식 (11), (12), (13)의  $N$ 은 실험 데이터 수,  $L_i'$ 이 분류기로 추측한 레이블 집합,  $L_i$ 은 실제 실험 데이터의 레이블 집합이라 정의한다. 첫 번째 성능 수치는 해밍 손실 (Hamming loss, H-Loss) 이다 [13]. 실제 레이블 집합에 속하지만 추측하지 못한 레이블의 개수와 추측한 레이블 집합에 속하지만, 실제 레이블 집합에 속하지 않은 레이블의 개수를 합쳐 평균을 내어 구한다. 식 (11)은 해밍 손실 수치를 표현한 것이다.

$$H-Loss = \frac{1}{N} \sum_{i=1}^N \frac{1}{q} |L_i' \triangle L_i| \dots (11)$$

해밍 손실 수치는 얼마나 오류 없이 레이블을 추측하는지, 그리고 실제 레이블을 제대로 측정하는지를 수치로 보여준다. 그러므로 해밍 손실 수치가 낮을수록 좋은 알고리즘이라 판단할 수 있다. 두 번째 성능 수치는 다중 레이블 정확성 (Multi-label accuracy, ML-ACC)이다 [13]. 추측한 레이블 집합과 실제 레이블 집합의 교집합 원소의 개수를 추측한 레이블 집합과 실제 레이블 집합의 한 집합 원소의 개수로 나누어 계산한다. 식 (12)은 다중 레이블 정확성을 표현한 것이다.

$$ML-ACC = \frac{1}{N} \sum_{i=1}^N \frac{|L_i' \cap L_i|}{|L_i' \cup L_i|} \dots (12)$$

다중 레이블 연관성 수치는 분류기가 오류를 최소화함과 동시에 얼마나 많은 레이블을 올바르게 추측해냈는지를 보여준다. 그러므로 다중 레이블 연관성 수치가 높을수록 좋은 알고리즘이라 판단할 수 있다. 식 (13)은 부분 정확성 (Subset accuracy, ST-ACC) 측정 수식이다 [13]. 이때 식 (13)은 추측한 레이블 집합이 실제 레이블 집합과 정확하게 일치하면 1, 아니면 0을 부여한 후, 부여한 값들을 모아 평균을 내어 성능 수치를 계산한다.

$$ST-ACC = \frac{1}{N} \sum_{i=1}^N [L_i' = L_i] \dots (13)$$

부분 정확성 수치는 오류 없이 얼마나 많은 개수를 정확하게 측정했는지를 보여준다. 그러므로 부분 정확성 수치가 높을수록 좋은 알고리즘이라고 판단할 수 있다.

4.3. Performance Comparison and Statistical Test

전체 30번 반복 실행한 결과를 통해 얻은 평균과 표준편차를 통해 각 알고리즘의 성능을 통계적으로 비교하였다 [19]. 분산 분석을 위해 프리드먼 검정 (Friedman test), 사후 검정은 본 페로니 검정(Bonferroni correction)을 사용했으며, 분산 분석의 동분산 가설 기각 기준 확률은 5%, 사후 검정의 임계값은 5% 기준인 2.241을 사용했다 [19].

4.4. Experimental Results

Table 2, 3, 4.은 30번 측정된 성능 수치의 평균과 표준편차를 나타낸 것이다. 대다수의 경우 본 논문이 제안한 알고리즘의 성능 수치가 높게 있음을 확인할 수 있다.

Table 2. Average and standard Deviation of H-Loss(↓)

Data set	MLNB-BR	MLNB-ECC	Proposed
Arts	.1071 ± .0036	.1087 ± .0038	<b>.0681 ± .0012</b>
Bibtex	.0809 ± .0021	.1412 ± .0047	<b>.0240 ± .0004</b>
Cal500	<b>.1400 ± .0030</b>	.1773 ± .0134	.1642 ± .0033
Emotions	.2315 ± .0168	.2278 ± .0165	<b>.2186 ± .0176</b>
Enron	.2122 ± .0079	.2117 ± .0067	<b>.0885 ± .0029</b>
Genbase	<b>.0071 ± .0019</b>	.0085 ± .0023	.0084 ± .0026
Health	.0724 ± .0030	.0748 ± .0032	<b>.0454 ± .0011</b>
L-Log	.3349 ± .0168	.4611 ± .0241	<b>.0199 ± .0013</b>
Medical	<b>.0192 ± .0008</b>	.0194 ± .0008	.0283 ± .0011
Scene	.1873 ± .0061	.1819 ± .0062	<b>.1445 ± .0061</b>
Slashdot	<b>.0407 ± .0013</b>	.0412 ± .0013	.0436 ± .0016
tmc2007	.1392 ± .0011	.1444 ± .0012	<b>.1053 ± .0008</b>
Yeast	.2684 ± .0082	.2816 ± .0118	<b>.2285 ± .0059</b>

Table 3. Average and standard Deviation of ML-ACC(↑)

Data set	MLNB-BR	MLNB-ECC	Proposed
Arts	.3293 ± .0092	.3391 ± .0099	<b>.4059 ± .0095</b>
Bibtex	.1924 ± .0060	.1710 ± .0063	<b>.2520 ± .0081</b>
Cal500	.1954 ± .0150	<b>.2179 ± .0158</b>	.2106 ± .0088
Emotions	.5510 ± .0299	.5577 ± .0278	<b>.5682 ± .0287</b>
Enron	.1986 ± .0091	.2018 ± .0093	<b>.2772 ± .0136</b>
Genbase	.8981 ± .0262	.8876 ± .0275	<b>.9117 ± .0262</b>
Health	.4770 ± .0101	.4808 ± .0107	<b>.5212 ± .0090</b>
L-Log	<b>.0375 ± .0015</b>	.0303 ± .0015	.0137 ± .0057
Medical	.3348 ± .0224	.3331 ± .0220	<b>.4195 ± .0216</b>
Scene	.5282 ± .0135	.5363 ± .0130	<b>.5961 ± .0165</b>
Slashdot	.4551 ± .0165	.4727 ± .0168	<b>.5533 ± .0146</b>
tmc2007	.3949 ± .0033	.3909 ± .0038	<b>.4461 ± .0028</b>
Yeast	.4280 ± .0129	.4295 ± .0162	<b>.4489 ± .0104</b>

Table 4. Average and standard Deviation of ST-ACC(↑)

Data set	MLNB-BR	MLNB-ECC	Proposed
Arts	.1623 ± .0144	.1756 ± .0136	<b>.2861 ± .0097</b>
Bibtex	.0633 ± .0053	.0654 ± .0058	<b>.0986 ± .0075</b>
Cal500	<b>.0000 ± .0000</b>	<b>.0000 ± .0000</b>	<b>.0000 ± .0000</b>
Emotions	.2624 ± .0419	.2918 ± .0445	<b>.3186 ± .0352</b>
Enron	.0006 ± .0012	.0010 ± .0016	<b>.0100 ± .0061</b>
Genbase	.8513 ± .0338	.8328 ± .0360	<b>.8742 ± .0330</b>
Health	.2289 ± .0134	.2370 ± .0133	<b>.3113 ± .0124</b>
L-Log	.0000 ± .0000	.0000 ± .0000	<b>.1401 ± .0138</b>
Medical	.2998 ± .0236	.2952 ± .0237	<b>.3768 ± .0246</b>
Scene	.2482 ± .0171	.2564 ± .0167	<b>.3872 ± .0457</b>
Slashdot	.3672 ± .0174	.3787 ± .0180	<b>.4840 ± .0159</b>
tmc2007	.0856 ± .0033	.0852 ± .0044	<b>.1295 ± .0038</b>
Yeast	.0985 ± .0112	.1129 ± .0148	<b>.1337 ± .0126</b>

Table 5. Result of friedman test from H-Loss

	MLNB-BR	MLNB-ECC	Proposed
Rank average	1.92308	2.61538	<b>1.46154</b>
$\sum \chi^2$	8.77	<i>p</i>	0.01247

Table 6. Result of friedman test from ML-ACC

	MLNB-BR	MLNB-ECC	Proposed
Rank average	2.53846	2.23077	<b>1.23077</b>
$\sum \chi^2$	12.15	<i>p</i>	0.00230

Table 7. Result of friedman test from ST-ACC

	MLNB-BR	MLNB-ECC	Proposed
Rank average	2.76923	2.23077	<b>1.00000</b>
$\sum \chi^2$	21.38	<i>p</i>	0.00002

Table 5, 6, 7.은 각 성능 수치에 순위를 매겨 프리드먼 검정을 시행한 결과이다. 각 Table 5, 6, 7.의 *p*는 각 성능 수치의 평균이 같을 확률을 나타낸다. 이 실험의 동분산 기각 기준 확률은 5% (0.05) 이므로 모든 성능 기준에 대해 각 알고리즘의 성능은 통계적으로 같을 수 없음을 확실히 보여준다.

이제 각 성능수치에 대해 각 알고리즘의 우열을 확실히 검증하기 위해 본 페로니 검정을 사용했다. 본 페로니 검정의 임계 수치는 다른 논문을 참조 [19]하여 계산하면  $2.241 \sqrt{\frac{3 \cdot 4}{6 \cdot 13}} = 0.8790$ 이다. 본 페르니 검정에 따르면, 특정 분류기의 평균 순위에 저 임계수치를 뺀 평균 순위 수치보다 더 높은 평균 순위를 가진 분류기는 특정 분류기보다 95% 확률로 성능이 높음을 통계적으로 검정할 수 있다. Fig. 1, 2, 3.은 각 수치의 평균 순위와 임계수치를 수직선 위에 나타내 성능을 비교한 것이다. 이때 Fig. 1, 2, 3.의 CD는 임계수치를 나타낸다.

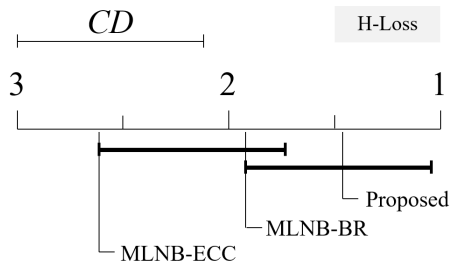


Fig. 1. Comparison of three classifiers using H-loss

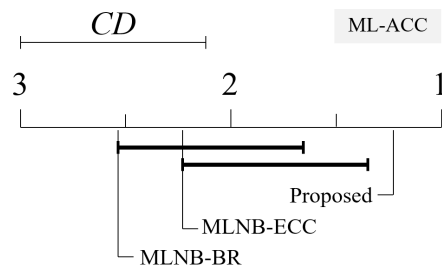


Fig. 2. Comparison of three classifiers using ML-ACC

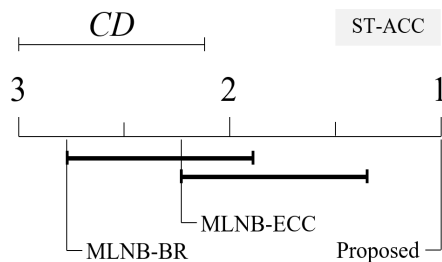


Fig. 3. Comparison of three classifiers using ST-ACC

Fig. 1.는 H-Loss의 관점에서 진행된 비교 결과를 보여준다. 본 논문에서 제안한 분류기의 평균 성능 순위가 MLNB-BR의 H-Loss 평균 순위의 임계수치 안쪽 범위에 있기 때문에, 본 논문의 분류기가 MLNB-BR보다 통계적으로 H-Loss 성능이 좋다고 확신할 수 없다. 하지만 본 논문 분류기의 H-Loss 평균 순위가 MLNB-ECC의 H-Loss 평균 순위의 임계수치 바깥에 있으므로 본 논문의 H-Loss 성능이 MLNB-ECC보다 통계적으로 성능이 좋음을 확인할 수 있다.

Fig. 2, 3.은 각각 ML-ACC의 관점과 ST-ACC의 관점에서 진행된 비교 결과를 나타낸 것이다. 두 관점 모두 본 논문의 제안 분류기 성능 순위가 MLNB-BR과 MLNB-ECC의 임계수치 바깥에 존재함을 확인할 수 있다. 이를 통해 본 논문의 제안 분류기의 ML-ACC, ST-ACC 성능이 타 비교 알고리즘들보다 통계적으로 좋은 성능을 가지고 있음을 확인할 수 있다.

결과적으로 대부분의 실험에서 기존 확장방식을 사용하여 확률을 추정하던 다중 레이블 나이브 베이즈 분류기보다 본 논문에서 제안한 사후확률 식을 사용한 다중 레이블 나이브 베이즈 분류기의 여러 성능이 더 좋음을 확인할 수 있다.

## V. Conclusions

본 논문은 각 레이블 간 연관성을 추측에 반영하는 사후확률 추정 식을 제안했고, 이를 다중 레이블 나이브 베이즈 분류기에 적용하였다. 그리고 본 논문에서 제안한 확률 추정 식은 레이블의 연관성을 확률 추정 식에 반영하여 높은 추측 성능을 끌어낼 수 있음을 확인했으며, 통계적인 기법을 통해 기존 확장 기법을 적용한 나이브 베이즈 분류기들보다 본 논문에서 제안한 수식을 사용한 나이브 베이즈 분류기가 더 좋은 성능을 끌어낼 수 있음을 확인했다.

이를 통해 특히 레이블 간의 연관성이 중요한 데이터를 사용하는 분야에서의 정확도 상승을 꾀할 수 있을 것으로 기대된다. 이후 후속 연구를 통해 본 연구에서 제안한 사후확률 추정 식을 나이브 베이즈 분류기가 아닌 다른 단일 레이블 분류기에 적용하여 더 나은 성능을 보일 수 있는지 시도할 것이다.

## REFERENCES

- [1] Lee, Jaedong, et al., "An approach for multi-label classification by directed acyclic graph with label correlation maximization," *Information Sciences*, Vol. 351, pp. 101-114, March 2016.
- [2] Sucar, L. Enrique, et al., "Multi-label classification with Bayesian network-based chain classifiers," *Pattern Recognition Letters*, Vol. 41, pp. 14-22, November 2014.
- [3] Zhang, Min-Ling, and Zhi-Hua Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern recognition*, Vol. 40, No. 7, pp. 2038-2048, 2007.
- [4] Zhang, Min-Ling, José M. Peña, and Victor Robles, "Feature selection for multi-label naive Bayes classification," *Information Sciences*, Vol. 179, No. 19, pp. 3218-3229, June 2009.
- [5] Godbole, Shantanu, and Sunita Sarawagi, "Discriminative methods for multi-labeled classification," *Pacific-Asia conference on knowledge discovery and data mining*, pp. 22-30, Berlin, Heidelberg, Germany, 2004.
- [6] Read, Jesse, et al., "Classifier chains for multi-label classification," *Machine learning*, Vol. 85, No. 3, pp. 333, 2011.
- [7] Ueda, Naonori, and Kazumi Saito, "Parametric mixture models for multi-labeled text," *Advances in neural information processing systems*, pp. 737-744, 2003.
- [8] I. Katakis et al., "Multilabel Text Classification for Automated Tag Suggestion," 2008 Discovery Challenge, Antwerp, Belgium, September, 2008.
- [9] D. Turnbull et al., "Semantic Annotation and Retrieval

- of Music and Sound Effects," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 16, No. 2, pp. 467-476, 2008.
- [10] Trohidis, Konstantinos, et al., "Multi-Label Classification of Music into Emotions," *ISMIR*, Vol. 8, pp. 325-330, 2008.
- [11] Elisseeff, André, and Jason Weston, "A kernel method for multi-labelled classification," *Advances in neural information processing systems*, pp. 681-687, 2002.
- [12] A. Cano et al., "LAIM discretization for multi-label data," *Information Science*, Vol. 330, pp. 370-384, 2016.
- [13] Zhang, Min-Ling, and Zhi-Hua Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, Vol. 26, No. 8, pp. 1819-1837, 2014.
- [14] Klimt, Bryan, and Yiming Yang, "The enron corpus: A new dataset for email classification research," *European Conference on Machine Learning*, pp. 217-226, Springer, Berlin, Heidelberg, 2004.
- [15] Boutell, Matthew R., et al., "Learning multi-label scene classification," *Pattern recognition*, Vol. 37, No. 9, pp. 1757-1771, 2004.
- [16] Barutcuoglu, Zafer, Robert E. Schapire, and Olga G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, Vol. 22, No. 7, pp. 830-836, 2006.
- [17] Joachims, Thorsten, "Text categorization with support vector machines: Learning with many relevant features," *European conference on machine learning*, pp. 137-142 Springer, Berlin, Heidelberg, 1998.
- [18] Read, Jesse, et al., "Classifier chains for multi-label classification," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 254-269, Springer, Berlin, Heidelberg, 2009.
- [19] Demšar, Janez, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, pp. 1-30. January 2006.
- [20] M. Zhang and L. Wu, "LIFT: Multi-label Learning with Label-Specific Features," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 37, No.1, pp. 107-120, 2015.
- [21] S. Diplaris et al., "Protein Classification with Multiple Algorithms," *Proc. 10th Panhellenic Conf. Inf.*, pp. 448-456, Volos, Greece, Nov 2005.
- [22] J. Pestian et al., "A shared task involving multi-label classification of clinical free text," *Proc. Work. BioNLP 2007*, pp. 97-104, Prague, Czech, June 2007.
- [23] A. Srivastava and B. Zane-Ulman, "Discovering recurring anomalies in text reports regarding complex space systems," *Proc. 2005 IEEE Aerospace Conf.*, pp. 3853-3862, Big Sky, USA, Mar 2005.
- [24] Baek, Yeong Tae, "Detection of Character Emotional Type Based on Classification of Emotional Words at Story," *Journal of the Korea Society of Computer and Information*, Vol. 18, No. 9, pp. 131-138, 2013.
- [25] Wang, Tinghuai, Jean-Yves Guillemaut, and John Collomosse, "Multi-label propagation for coherent video segmentation and artistic stylization," *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 3005-3008, 2010.
- [26] Persing, Isaac, and Vincent Ng, "Modeling thesis clarity in student essays," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 260-269, 2013.
- [27] Park, Jong-Beom, "Development of the Poker Game Achievement Engine for Artificial Intelligence," *Journal of the Korea Society of Computer and Information*, Vol. 14, No. 11, pp. 41-52, 2009.

## Authors



Hae-Cheon Kim received the B.S. degrees in Computer Science and Engineering from Chang-Ang University, Korea, in 2018. He is studying in graduate school of computer engineering at Chang-Ang University. He is interested in optimization of algorithms

about artificial Intelligence and graph theory.



Jaesung Lee is currently an assistant professor in the School of Computer Science and Engineering, Chung-Ang Univ. in Seoul, Korea. Prior to coming to CAU, he did his postdoc, Ph.D., M.S. and B.S. at Chung-Ang Univ., Korea. His research

interest includes advanced machine learning algorithms with innovative applications to music emotion recognition, educational data mining, affective computing, and robot interaction.