다중 레이블 나이브 베이지안 분류기의 정확도 개선 연구

김해천[O], 이재성[*]
[O*]중앙대학교 컴퓨터공학부
e-mail: code.bug.station@gmail.com[O], jslee.cau@gmail.com[*]

Improving Accuracy of Multi-label Naive Bayes Classifier

Hae-Choen Kim[O], Jae-Sung Lee[*]
[O*]OSchool of Computer Science and Engineering, Chung-Ang University

● 요 약 ●

다중 레이블 분류 문제는 다중 레이블 데이터를 입력받았을 때 연관된 다수의 레이블을 추측하는 문제이다. 본 논문에서는 다중 레이블 분류 문제의 기법 중 하나인 나이브 베이지안 분류기에 레이블 의존성을 계산하여 결과에 반영한 결과 다중 레이블 분류 문제의 성능이 개선됨을 확인하였다.

키워드: 다중 레이블 분류(Multi-label Classification), 나이브 베이즈(Naive Bayes), 레이블 의존성(Label Dependency)

I. 소개

다중 레이블 분류 문제는 다중 레이블 데이터가 주어졌을 때 연관된 다수의 레이블들을 추측하는 문제로, 최근 여러 방면으로 연구되고 있다 [1]. 여기서 다중 레이블 데이터란 특정 패턴이 하나 이상의 레이블과 연관성을 가지는 데이터를 뜻한다.

여러 다중 레이블 분류기 중 다중 레이블 나이브 베이지안(Multi-Label Naive Bayesian) 분류기가 널리 쓰이고 있다 [2]. 하지만, 이 분류기는 레이블 간의 상호관계를 결과에 반영할 수 없는 문제점을 가지고 있다. 이 문제를 보완하기 위해 본 연구에서는 레이블 사이의 의존성을 반영하는 나이브 베이지안 분류기를 제안하고자 한다. 그리고 실험을 통하여 본 논문에서 제안한 분류 기법과 기존에 널리 쓰이는 다중 레이블 분류기와의 결과를 비교하여 본 논문에서 제안한 방식이 더 높은 정확도를 가지고 있음을 보이고자 한다.

II. 관련 연구

다중 레이블 분류 문제는 데이터 $[x_1, x_2, \dots, x_d]$을 입력받았을 때 연관된 레이블 $\{l_1, l_2, \dots, l_q\}$을 추측하는 문제이다. 다중 레이블 분류기는 이 문제를 풀기 위한 알고리즘을 뜻한다.

최근 다중 레이블 분류 문제를 해결하기 위한 접근 기법으로 각각 레이블마다 단일 레이블 분류기를 생성하여 각자 분류하여 취합하는 이진 연관성(Binary Relevance, BR) 기법이 있다. 대표적으로 다중 레이블 나이브 베이지안(MLBRNB) [3]가 있다. 둘째로 가까운 k개의 샘플의 값을 이용하여 모든 레이블의 존재를 판단하는 다중 레이블 k-최근접 이웃(Multi Label k- Nearest Neighbor, ML-kNN)이 있다 [4].

III. 제안하는 방법론

기호 $l_k^{r_k}$에서 레이블 l_k가 다중 레이블 집합에 속하면 r_k을 1로, 아니면 0으로 나타내기로 한다. 본 연구에서는 단일 레이블 나이브 베이지안 분류기의 사후확률 $P(L|x_1x_2\dots)$의 L을 $l_1^{r_1}l_2^{r_2}\dots l_q^{r_q}$로 바꿔 다중 레이블 문제로 치환하는 방식을 택했다. 그리고 각 데이터들끼리 독립이고 각 레이블들도 서로 조건부 독립이라 가정하고 사후확률 수식을 풀어 식 (1)을 유도했다(이때 t는 임의의 레이블 번호).

$$\underset{r_1r_2\dots}{\operatorname{argmax}} \prod_{n=1}^{d} P(x_n|l_t^{r_t}) \prod_{m=1}^{q} P(l_m^{r_m}|l_t^{r_t}) P(l_t^{r_t}) \qquad \textbf{(1)}$$

본 논문에서 제안하는 알고리즘은 위 식(1)에 가능한 모든 레이블 조합($l_1^{r_1}l_2^{r_2}\dots l_q^{r_q}$)에 모든 레이블 t을 대입하여 식(1)의 값을 모두 계산하고 평균을 내어 결과가 가장 큰 조합을 선택하는 것이다. 이론상 조합 가능한 레이블의 개수는 2^q로 매우 크지만 실세계 데이터에 나타나는 레이블 조합의 개수는 매우 한정적이고, 레이블끼리의 의존성 수치는 Memoization을 통해 계산 효율성을 높일 수 있다.

V. 실험 결과

제안하는 알고리즘의 성능을 검증하기 위해 실세계 데이터인 Bugs2664, Emotions, Yeast [5,6]으로 본 논문이 제안하는 알고리즘

과 MLBRNB, 그리고 ML-kNN 방식을 비교할 것이다. [표 1]은 데이터의 특징을 서술한 것이다.

교차검증을 위해 각 데이터를 무작위로 선택하여 80%의 학습 데이터와 20%의 평가 데이터로 구분하였다. 그리고 학습 데이터를 이용하여 MLBRNB과 ML-kNN (k=10), 그리고 본 논문에서 제시한 알고리즘을 학습시키고, 20%의 평가데이터를 이용하여 정확도를 구하였다. 여기서 정확도를 평가하는 기준은 Multi-label Accuracy를 사용하였다. [2]

데이터를 나누고 학습시키고 평가하는 과정을 50번 반복하여 정확도를 평가하고, 그 결과들로 평균과 편차를 구하였다. [표 2]은 그 결과 값을 보여준다. 표의 수치가 높을수록 실제 레이블들과 예측 레이블들이 정확하게 일치함을 나타낸다. [표 1]의 수치를 통해 이 논문에서 제시한 분류기의 성능이 기존의 MLBRNB나 ML-kNN보다 높아졌음을 알 수 있다.

Table 1. 데이터 명세

Data set	Domain	Patterns	Features	Labels
Bugs2664	Tag	2664	137	40
Emotions	Emotion	593	72	6
Yeast	Biology	2417	103	14

Table 2. Multi-Label Accuracy 측정값의 평균±분산 결과

Data set	MLBRNB	ML-kNN	Proposed
Bugs2664	**0.1053** ±0.0057	**0.0041** ±0.0023	**0.1372** ±0.0082
Emotions	**0.5474** ±0.0275	**0.5175** ±0.0261	**0.5646** ±0.0267
Yeast	**0.4266** ±0.0140	**0.4298** ±0.0148	**0.4486** ±0.0126

V. 결론

본 논문에서는 다중 레이블 분류 기법 중 나이브 베이지안 분류 기법에 레이블 의존성을 반영하는 분류기를 제안하였다. 그리고 레이블 의존성을 분류기에 적용하고 실험을 해본 결과, 기존 분류 기법보다 다중 레이블 정확도가 높은 결과를 얻을 수 있음을 확인하였다.

REFERENCES

[1] Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." International Journal of Data Warehousing and Mining Vol. 3, No. 3, pp.64-67. 2006.

[2] 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.

[3] 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, 2009.

[4] 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.

[5] D.-W. Kim, "Korean Situation-Context-Emotion based Music Recommendation System," National Research Foundation of Korea, NRF-2010-0012885, 2013.

[6] Mulan: A Java Library for Multi-Label Learning, “http://mulan.sourceforge.net/dataset.html”