

변수 간 상호 의존성을 반영한 개선형 나이브 베이즈 분류기

이정훈, 김철호, 이병화, 백옥기
한국전자통신연구원

jhlee0914@etri.re.kr, kimcheolho@etri.re.kr, byounghwa.lee@etri.re.kr, ok.baek@etri.re.kr

Two-dependence augmented weighted naïve Bayes classifiers based on mutual dependencies between variables

Jung-Hoon Lee, Cheol Ho Kim, Byounghwa Lee, OK Baek
Electronics and Telecommunications Research Institute

요약

NB는 클래스에 대한 변수 간의 상호 독립성을 가정하는 단순한 구조를 취하기 때문에 데이터에 내재한 변수 간의 복잡한 상호 의존성을 반영하기에 한계가 있다. 본 논문에서는 변수 간의 상호 의존성을 보다 세밀하게 반영할 수 있는 two-dependence augmented weighted naïve Bayes(TDAWM)을 제안한다. TDAWM은 NB를 기반으로 하되 한 변수 노드가 최대 두 개의 다른 노드에 대한 연결을 가질 수 있도록 허용하고, 각 변수에 변수와 클래스, 변수 사이의 상호정보를 활용한 가중치를 적용하였다. 다양한 데이터에 대한 테스트 결과 TDAWM이 기존의 NB 계열 분류기들에 비해 개선된 성능을 보임을 검증하였다.

I. 서론

나이브 베이즈(naïve Bayes, NB) 분류기는 변수 간의 상호 독립성을 가정하고 있지만, 그러한 조건이 만족되지 않는 데이터에도 우수한 성능을 보인다. [1-2] NB의 성능을 개선하는 방법으로는 각 변수에 서로 다른 가중치를 적용하는 weighted NB(WNB)[3-4], 변수마다 최대 하나의 엷지를 다른 변수 방향으로 연결하는 tree-augmented NB(TAN)[5], 변수 중 하나를 다른 변수들과 모두 연결하는 최상위 부모(super parent)로 설정하고 최상위 부모가 다른 여러 모델들의 양상부를 취하는 averaged one-dependence estimator(AODE)[6] 등이 있다. WNB는 NB의 구조를 그대로 유지하기 때문에 모델의 표현력 증대에 제한적인 면이 있고, TAN은 변수 간 연결이 하나로 제한되어 복잡한 상호의존성을 담아내는 데 한계가 있으며, AODE는 양상부 방법이기에 변수가 많을 경우 모델의 복잡도가 지나치게 증가하여 최적화에 어려움이 있다는 단점이 있다.

본 논문에서는 데이터에 내재한 변수 간의 복잡한 상호 의존성을 보다 효과적으로 반영하기 위해 NB를 기반으로 하여 한 변수로부터 다른 변수로 향하는 연결을 최대 두 개만 허용하고, 변수마다 상호정보(mutual information) 기반의 가중치를 적용하여 모델의 복잡도가 크게 증가하지 않으면서도 예측 성능이 우수한 two-dependence augmented weighted NB(TDAWM) 분류기를 제안한다.

II. 제안 방법

본 논문에서 제안하는 TDAWM은 클래스에 대한 변수 간의 상호 독립성을 가정한 NB의 구조를 그 출발점으로 한다. 여기에 변수 간의 의존적 관계를 반영하기 위해 각

변수 사이의 연결을 허용하는 방식으로 구조적 확장을 수행하고, 이후 각 변수에 변수 간 상호 정보 기반의 가중치를 적용함으로써 모델의 표현력을 개선한다.

구조적 확장은 hill-climbing 기법을 통해 모델의 데이터에 대한 표현력을 계량화한 Bayesian-Dirichlet equivalent uniform(BDeu) 지표를 최대화함으로써 이루어진다. 이때 클래스를 제외한 변수 간의 연결에 대해서만 최적화를 진행하며, 보다 효율적인 최적화를 위해 변수 간 연결을 최대 두개까지 허용하여 계산 복잡도를 $O(n^2)$ 로 제한한다[7].

구조 학습 후 두 변수 사이의 상호정보를 기반으로 하여 각 변수의 가중치를 계산한다. 가중치는 하나의 변수가 클래스 변수에 대해 가지는 관련도(relevance)와 나머지 변수들에 대해 가지는 중복도(redundancy)를 계량화한 후, 관련도를 최대화하고 중복도를 최소화하는 minimum-redundancy-maximum-relevance(mRMR) 방식을 기반으로 추출한다[8]. i 번째 변수를 X_i 라고 할 때, X_i 의 가중치 w_i 는 아래와 같이 구할 수 있다.

$$w_i = \text{sigmoid}(\text{relevance} - \text{redundancy}) \\ = \text{sigmoid}(D(X_i) - R(X_i))$$

여기서 $D(X_i)$ 는 X_i 와 클래스 변수와의 상호정보를, $R(X_i)$ 는 X_i 와 다른 변수들과의 평균적인 상호정보를 나타내며 두 항은 모두 정규화된 값을 취한다.

III. 실험 및 분석

이제 제안한 방법에 따라 구현된 TDAWM 분류기의 성능을 검증하고 기존의 모델과 비교해보고자 한다. 학습과 검증을 위해 UCI ML Repository [9]에서 다섯 개의 데이터셋을 선정하였으며, 데이터셋에 대한 정보를 표 1에 요약하였다.

Name	Features	Instances	Classes
Cardiotocography	21	2126	3
Robot-24	24	5456	4
Satellite	36	6435	6
Splice	60	3190	3
Mice	77	1080	8

표 1. 실험 데이터셋

테스트마다 데이터셋을 임의로 75%의 학습 데이터와 25%의 검증 데이터로 나누었고, 이 과정을 5회 반복한 후 평균을 취하여 예측 결과를 도출하였다. TDAWN의 분류 예측 정확도를 표 2에 정리하였는데, 성능의 비교를 위해 NB, TAN, TDAN, WNB의 결과를 함께 명시하였다. TDAN은 TDAWN에서 가중치를 적용하지 않은 모델을 의미한다.

	NB	TAN	TDAN	WNB	TDAWN
Cardiotocography	83.20	88.12	89.06	85.19	89.66
Robot-24	82.92	90.37	92.17	86.06	93.58
Satellite	80.72	87.31	87.76	80.82	88.14
Splice	95.31	94.96	95.19	96.02	96.89
Mice	80.81	87.85	95.93	83.93	96.44

표 2. 성능비교표

검증 결과로부터 본 논문에서 제안한 TDAWN이 비교 알고리즘들에 비해 향상된 예측 성능을 가짐을 알 수 있다. 이러한 성능의 향상은 일차적으로 TDAWN이 기존의 NB나 TAN에 비해 확장된 구조적 복잡성을 통해 변수 간의 의존성을 보다 세밀하게 반영하기 때문이다. 이에 더하여 변수 가중치를 통해 변수와 클래스 및 변수 사이의 상관관계가 반영되기 때문에 보다 높은 예측 성능을 도출할 수 있다.

이외에도 TDAWN은 구조 학습과 변수 가중치 방법을 개별적으로 적용할 때보다 안정적인 성능 개선이 가능하다. 표 2의 Satellite 데이터에 대한 성능을 보면 구조 학습 방법은 효과적으로 NB의 성능을 개선하지만 변수 가중치 방법은 효과가 미미하다. 반대로 Splice 데이터에서는 변수 가중치 방법은 효과적인 반면 구조 학습 방법은 오히려 약간의 성능 하락을 가져온다. 이에 반해 TDAWN은 모든 데이터에서 효과적으로 NB의 예측 성능을 개선할 수 있다.

IV. 결론

본 논문에서는 데이터셋에 내재한 복잡한 상호관계를 반영할 수 있는 개선된 NB 모델인 TDAWN을 제안하고 UCI 데이터셋에 대해 예측 성능의 우수성을 검증하였다. TDAWN은 한 변수로부터 다른 변수로 향하는 연결을 최대 두 개까지 허용하고, 상호정보에 기반한 변수 가중치를 적용하여 모델이 담아낼 수 있는 확률분포의 복잡도를 증가시켰다. 구현 결과, 기존의 NB 계열의 모델들에 비해 우수한 분류 성능을 달성하여 본 논문의 접근 방법이 효과적임을 입증하였다.

TDAWN은 변수 노드와 노드 간의 연결을 직접적으로 확인할 수 있는 투명한 모델(white-box model)로서 결과에 대한 해석이 용이하여 설명 가능한 모델(explainable AI)로 활용될 수 있다. 또한 구조적인 확장과 가중치의 부여가 명확한 규칙에 기반하고 있으므로 변수 환경이 변화하는 동적인 환경에서 학습 모델을 점진적으로 성장시킬 수 있는 가능성도 내포하고 있어 폭넓은 응용이 기대할 수 있다.

ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음. [22ZS1100, 자율성장형 복합인공지능 원천기술 연구]

참고 문헌

- [1] P. Domingos, M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Mach. Learn.*, Vol. 29, pp. 103-130, 1997.
- [2] C.H. Lee, A gradient approach for value weighted classification learning in naïve Bayes, *Knowledge-Based Syst.*, Vol. 85, pp. 71-79, 2015.
- [3] L. Jiang, L. Zhang, C. Li, J. Wu, "A correlation-based feature weighting filter for naïve Bayes," *IEEE Trans. Knowl. Data Eng.*, Vol. 31, pp. 201-213, 2019.
- [4] H. Zhang, L. Jiang, L. Yu, "Attribute and instance weighted naïve Bayes," *Pattern Recognit.*, Vol. 111, 2021.
- [5] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, Vol. 29, pp. 131-163, 1997.
- [6] G. I. Webb, J. R. Boughton, Z. Wang, "Not so naïve Bayes: aggregating one-dependence estimators," *Mach. Learn.*, Vol. 58, pp. 5-24, 2005.
- [7] A. Perez, P. Larranaga, I. Inza, "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naïve Bayes," *Intern. J. Approx. Reasoning*, Vol. 43, pp. 1-25, 2006.
- [8] P. Maji, S. Paul, "Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data," *Int. J. Approx. Reason.*, Vol. 52, pp. 408-426, 2011.
- [9] D. Dua, C. Graff, UCI machine learning repository, University of California, Irvine, School of Information and Computer Sciences.