나이브베이즈분류기의 정확도 향상을 위한 자질변수통합

허민오° 김병희 황규백 장병탁

서울대학교 컴퓨터공학부

{moheo°, bhkim, kbhwang}@bi.snu.ac.kr, btzhang@cse.snu.ac.kr

Combining Feature Variables for Improving the Accuracy of Naïve Bayes Classifiers

Min-Oh Heo°, Byoung-Hee Kim, Kyu-Baek Hwang, and Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

나이브베이즈분류기(naïve Bayes classifier)는 학습, 적용 및 계산자원 이용의 측면에서 매우 효율적인 모델이다. 또한, 그 분류 성능 역시 다른 기법에 비해 크게 떨어지지 않음이 다양한 실험을 통해 보여져 왔다. 특히, 데이터를 생성한 실제 확률분포를 나이브베이즈분류기가 정확하게 표현할 수 있는 경우에는 최대의 효과를 볼 수 있다. 하지만, 실제 확률분포에 존재하는 조건부독립성(conditional independence)이 나이브베이즈분류기의 구조와 일치하지 않는 경우에는 성능이 하락할 수 있다. 보다 구체적으로, 각 자질변수(feature variable)들 사이에 확률적 의존관계(probabilistic dependency)가 존재하는 경우 성능 하락은 심화된다. 본 논문에서는 이러한 나이브베이즈분류기의 약점을 효율적으로 해결할 수 있는 자질변수의 통합기법을 제시한다. 자질변수의 통합은 각 변수들 사이의 관계를 명시적으로 표현해 주는 방법이며, 특히 상호정보량(mutual information)에 기반한 통합 변수의 선정이 성능 향상에 크게 기여함을 실험을 통해 보인다.

1. 서론

나이브베이즈분류기(naïve Bayes classifier)는 클래스변수(class variable)의 값이 주어진 경우에 모든 자질(feature)들은 조건부독립임(conditionally independent)을 가정하는 모델이다. 이는 그림 1과 같이 표현된다. 그림 1의 베이지안망(Bayesian network) 구조는 클래스변수 C가 주어진 경우 X_i, X_j ($1 \le i, j \le n, i \ne j$)는 서로 독립임을 나타내고 있다. 이러한 나이브베이즈분류기는 그 학습, 적용이 매우 효율적이다. 모델을 구성하는 파라미터는 확률분포 $P(C)$와 $P(X_i|C)$에 대한 파라미터로 한정된다. 학습된 모델의 적용 역시 (수식 1)과 같이 효율적으로 행해질 수 있다. (수식 1)에서 자질변수들에 대한 확률분포(marginal probability distribution)는 결과와 관계가 없으며, 새로운 데이터의 분류는 간단한 계산으로 이루어진다.

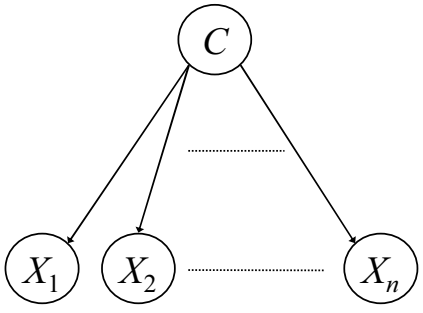

그림 1. 나이브베이즈분류기

$$P(C \mid X_1, X_2, ..., X_n) = \frac{P(C)P(X_1, X_2, ..., X_n \mid C)}{P(X_1, X_2, ..., X_n)} = \frac{P(C)\prod_{i=1}^{n} P(X_i \mid C)}{P(X_1, X_2, ..., X_n)}$$

(수식 1)

이러한 나이브베이즈분류기는 실제 데이터를 생성한 확률분포가 가정된 조건부독립성(conditional independence) 가정을 만족한다면 최적의 성능을 발휘하게 된다. 나이브베이즈분류기의 분류 성능은 다양한 실험을 통해, 그리고 이론적으로 입증되어 왔다 [1]. 하지만, 실세계의 문제들은 그러한 나이브베이즈 가정(naïve Bayes assumption)을 따르지 않는 경우가 많다. 구체적으로, 각 자질변수들이 조건부독립이 아닌 경우 성능의 하락이 예상된다. 또한, 나이브베이즈분류기의 표현력은 각 변수들이 이진값을 가지는 경우에는 선형분류기(linear classifier)와 동일함이 증명되어 있다 [2].

나이브베이즈분류기의 구조적 제약을 완화하려는 시도는 꾸준히 제시되어 왔다. 직관적으로, 강력한 조건부독립성을 완화하기 위해 각 자질변수들 사이의 연관관계를 찾고 간선을 추가하는 방법들이 제시되어 왔다 [3]. 본 논문에서는 자질변수들의 합성을 통해 나이브베이즈분류기의 성능향상을 꾀하는 새로운 기법을 제시한다. 이러한 기법은 구조의 학습 과정을 거치지 않으면서 각 변수

들 사이의 관계를 표현할 수 있다. 자질변수의 합성을 위한 기준으로는 각 변수들 사이의 상호정보량(mutual information)을 이용한다. 본 논문은 다음과 같이 구성되어 있다. 2절에서는 본 논문에서 제시하는 자질합성 기법을 기술한다. 그리고 3절에서는 유전자발현데이터를 이용해서 제시한 기법의 성능을 실험적으로 평가하며, 마지막으로 4절에서 결론을 제시한다.

2. 자질변수 통합

본 논문에서는 이산변수(discrete variable)들로 구성되어 있는 문제를 다룬다. 이산변수 X가 가질 수 있는 값의 집합을 $\mathbf{d}X$라고 하자. 두 이산 변수 X, Y가 있는 경우, 이러한 두 변수의 통합은 다음과 같이 이루어진다. 새로운 변수 XY는 $\mathbf{d}X$의 모든 원소와 $\mathbf{d}Y$의 모든 원소들의 순서쌍을 값으로 가진다. 상기의 변수 통합을 통해 두 변수 사이의 임의의 관계를 모두 표현할 수 있으며, 이는 필요한 파라미터 수의 증가를 그 비용으로 한다. 극단적인 경우 모든 자질변수들을 통합하면 가능한 결합확률분포를 모두 표현할 수 있다.[1]

본 논문에서는 두 변수 사이의 통합만을 고려한다. 문제를 구성하는 자질변수의 개수가 n인 경우, 통합의 대상이 되는 순서쌍의 개수는 $O(n^2)$으로 제한된다. 이러한 통합을 통해 늘어나는 파라미터의 개수는 통합되는 두 변수 사이에 간선을 하나 추가함으로써 늘어나는 파라미터의 개수와 같다. 이는 그림 2와 같이 보여진다.

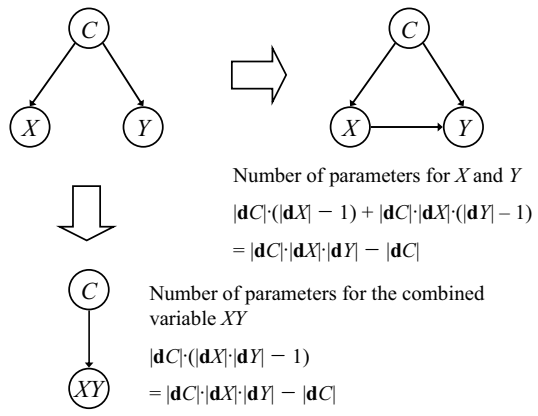

그림 2. 자질변수 통합과 간선 추가의 비교. 두 방법 모두 증가하는 파라미터의 개수는 동일하다.

모든 자질변수 쌍들을 통합하는 것은 비효율적이며, 두 변수의 결합이 분류에 영향을 주지 않는다면, 이는 파라미터의 개수만 증가시킴으로써 주어진 데이터가 한정된 경우 결국 잡음(noise)으로 작용하게 될 가능성이 커진다. 따라서, 통합할 변수 쌍들을 선정하는 것이 중요하다. 본 논문에서는 두 변수들 사이의 확률적 의존관계에 대한 척도인 상호정보량(mutual information)을 기준으로 삼는다. 두 자질변수 X, Y 사이의 상호정보량은 (수식 2)와 같이 계산된다.

$$I(X;Y) = \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \qquad \text{(수식 2)}$$

두 변수 사이의 상호정보량은 두 변수가 서로 독립인 경우 0이 되며, 한 변수가 다른 한 변수에 대한 함수로 표현되는 경우 최대값을 가지게 된다. 기준의 적용은 상호정보량이 큰 변수 쌍들을 통합하는 방법과 작은 변수 쌍들을 통합하는 방법이 있다. 상호정보량이 큰 변수들의 통합은 X, Y가 서로 확률적으로 의존인 경우 간선을 추가하는 방법과 점근적으로 같은 결과를 가져오는 방법이다. 하지만 이러한 방법의 경우, 비슷한 변수들을 통합함으로써 불필요한 정보를 모델에 포함시킬 가능성이 있다. 이에 반해 두번째 방법은 서로 다른 양상을 보이는 변수들을 통합함으로써, 기존의 나이브베이즈분류기로는 표현할 수 없었던 관계를 표현하게 된다.

3. 실험

3.1 실험데이터

본 논문에서는 제시된 기법을 평가하기 위해 실제 마이크로어레이(microarray) 데이터를 이용하였다. 마이크로어레이는 샘플에서의 수천~수만에 달하는 유전자 발현량을 동시에 측정할 수 있는 기술이다 [4]. 이용된 마이크로어레이 데이터는 120개의 샘플로 구성된 백혈병 환자 데이터이다 [5]. 이들중 60개의 샘플은 특정 약을 처리하기 전의 샘플이며, 나머지 60개의 샘플은 특정 약을 처리한 이후의 샘플이다. 문제는 유전자발현양상에 기반하여 약물 처리 전과 후를 구분하는 것이다. 원래의 데이터는 다음과 같이 전처리되었다. 우선, 모든 유전자발현값들은 각 샘플에서의 중간값(median)에 기반하여 이진화되었다. 이후 12,600개의 유전자 중에서 분류와 관련이 깊은 30개의 유전자를 상호정보량에 기반하여 선정하였다. 결국 최종 데이터는 30개의 자질변수와 120개의 샘플로 구성된다.

3.2 실험결과

본 논문에서 제시한 방법에 의한 분류 성능의 변화는 그림 3 에 제시되어 있다. 그림 3 의 x 축은 통합된 자질변수 쌍의 개수를 나타내며 y 축은 10-fold cross validation 으로 평가된 분류정확도이다. 실선은 상호정보량이 낮은 순으로 자질변수를 통합한 경우이며, 점선은 상호정보량이 높은 순으로 자질변수를 통합한 경우를 보이고 있다. 결과를 보면 상호정보량이 높은 순으로 변수를 통합하는 것은 성능 저하를 가져옴을 알 수 있다. 이는 상호정보량이 비슷한 변수들은 분류를 행하는데 있어서 서로 중복된 정보를 가지고 있을 가능성이 크며, 이러한 변수들을 통합하는 것은 파라미터의 수를 늘림으로

[1] 이는 물론 변수의 개수가 조금만 많아도 비현실적이 된다.

써 불필요한 잡음을 생성시킬 가능성이 커지기 때문으로 생각된다. 한편, 상호정보량이 낮은 변수들의 통합은 두 변수들이 통합되어 분류를 하는데 필요한 정보량을 높임으로서 성능 향상에 크게 도움이 됨을 알 수 있다.

그림 4, 5 는 본 논문에서 제안한 자질변수의 통합이 결정트리 및 신경망에 미치는 효과를 보이고 있다. 결정

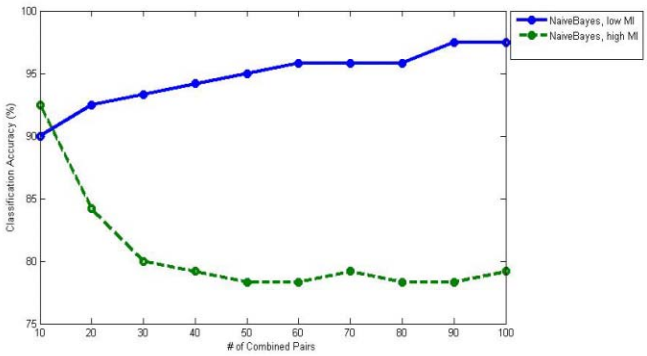

그림 3. 자질변수 통합에 의한 나이브베이즈분류기의 성능 변화. 상호정보량이 낮은 순서대로 자질변수들을 통합할 때 그 성능이 증가함을 알 수 있다 (실선).

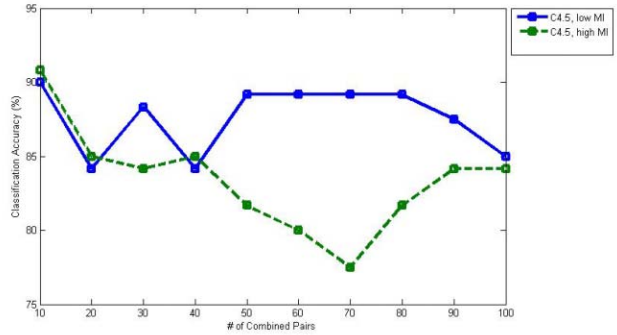

그림 4. 자질변수의 통합이 결정트리(C4.5)에 미치는 영향.

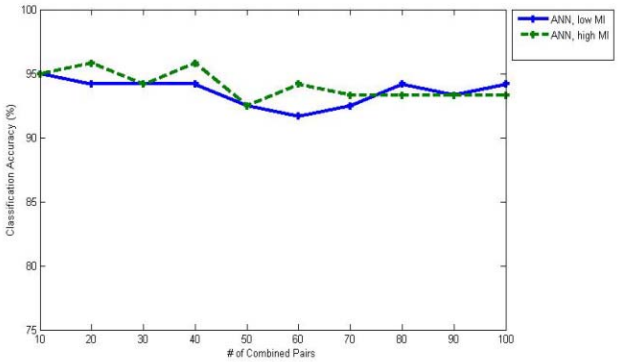

그림 5. 자질변수의 통합이 신경망(multilayer perceptron)에 미치는 영향.

트리의 경우 비선형 분류 문제의 해결에 적합한 모델로 자질변수의 통합이 성능의 하락을 가져오기 쉬움을 보이고 있다. 이는 자질변수의 통합이 나이브베이즈분류기의 표현력의 한계를 극복하기 위한 기법이라는 점에서 이와는 표현력이 다른 결정트리 모델의 경우 성능 향상을 기대하기 어려움을 보인다고 할 수 있다. 신경망(multilayer perceptron)의 경우에는 은닉노드를 통해서 각 변수들 사이의 관계를 표현할 수 있으므로 자질 변수의 통합이 성능에 영향을 주지 않음을 알 수 있다.

자질변수 통합이 나이브베이즈분류기의 성능을 크게 향상시킴은 그림 3, 4, 5 의 비교를 통해서 확인할 수 있다. 100 개의 자질변수 쌍을 상호정보량이 낮은 순으로 추가했을 때의 나이브베이즈분류기는 97.5%의 분류 정확도를 보임으로써, 다른 어떤 모델들보다도 뛰어난 성능을 보였다.

4. 결론

본 논문에서는 효율적인 분류기인 나이브베이즈분류기의 성능을 간단하게 향상시킬 수 있는 기법인 자질변수의 통합을 제시하였다. 통합을 위한 자질들의 선정에는 상호정보량을 이용하였으며, 상호정보량이 낮은 순서대로 변수들을 통합하는 것이 나이브베이즈분류기의 한계점인 표현력을 극복할 수 있는 좋은 방법임을 실험을 통해 보였다. 특히, 자질변수의 통합은 실제 나이브베이즈분류기 구조에 간선을 하나 추가하는 것과 같은 복잡도의 증가를 가져오는 반면, 구조의 학습 과정 등을 필요로 하지 않는 효율적인 방법이다.

감사의 글

이 논문은 교육부 BK21 사업, 과학기술부 국가지정연구실사업(NRL)에 의하여 지원되었음. 김병희는 서울과학장학생 사업에 의해 지원받았음.

참고 문헌

[1] Domingos, P. and Pazzani, M., On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning*, vol. 29, pp. 103-130, 1997.

[2] Ling, C. X. and Zhang, H., The representational power of discrete Bayesian networks, *Journal of Machine Learning Research*, vol. 3, pp. 709-721, 2002.

[3] Friedman, N., Geiger, D., and Goldszmidt, M., Bayesian network classifiers, *Machine Learning*, vol. 29, pp. 131-163, 1997.

[4] Knudsen, S., *A Biologist's Guide to Analysis of DNA Microarray Data*, John Wiley & Sons, New York, NY, 2002.

[5] Cheok, M. H., Yang, W., Pui, C.-H., Downing, J. R., Cheng, C., Naeve, C. W., Relling, M. V., and Evans, W. E., Treatment-specific changes in gene expression discriminate *in vivo* drug response in human leukemia cells, *Nature Genetics*, vol. 34, no. 1, pp. 85-90, 2003.