의료데이터마이닝에서 클러스터링 기반의 나이브 베이지안 학습

한송이° 정용규
을지대학교 의료전산학전공
ygjung@eulji.ac.kr

A Naive Bayesian Learning of Clustering for Medical Datamining

Song-yi Han° Young-Gyu Jung
Eulji University, Dept. of Medical Information Technology

요 약

병원정보시스템의 전세계적인 보급과 데이터웨어하우스의 도입으로 인해서 병원내의 의료데이터가 기하급수적인 증가추세를 보이고 있다. 환자에 대한 임상적인 특징을 다수 포함하고 있는 의료데이터는 유용한 임상지식의 보고로서 그 가치가 매우 유용하다. 따라서 데이터에 숨겨진 지식을 발견하여 구조화시킴으로써 새로운 지식을 창조하는 데이터마이닝은 임상부분에 적합한 기술이라 말할 수 있다. 본 연구에서는 급성염증을 가진 환자들의 의료데이터를 기반으로 특징을 추출하고, 추출된 특징을 바탕으로 병명을 판단하기 위한 학습을 수행한다. 학습 방법은 클러스터링을 이용한 나이브 베이지안으로 진행한다. 기존의 나이브 베이지안 학습은 대량의 데이터를 처리하는데 효과적이며 성능 또한 우수하지만, 속성별 독립을 가정하기 때문에 의료데이터를 분석에는 잘 사용되지 않는다. 따라서 높은 신뢰도를 구현하기 위해 나이브 베이지안 학습 전에 클러스터링을 선행하여, 기존 데이터에 클러스터링 클래스를 추가한다. 이를 통해 급성염증의 증상을 보이는 환자데이터를 바탕으로 자동적으로 방광염과 결석으로 인한 신장염을 효과적으로 진단해낸다.

키워드: 클러스터링, 클러스터링 클래스, 나이브 베이지안, 의료데이터마이닝

1. 서 론

의료데이터가 담고 있는 환자의 수많은 임상적 특징은 결과의 인과 관계를 설명하는 것으로 단순 조직화되어 있는 데이터라기보다는 환자의 진료과정의 부산물로서 인정된다.[1] 따라서 의료데이터는 질적으로 가치 있는 것이며, 의사들이 효율적인 치료패턴을 찾을 수 있게 도와줌으로 인해서 이에 근거한 진료의 방향을 결정하고, 의료의 질적 향상을 도모할 수 있게 한다.

본 논문에서는 실제 급성염증(Acute Inflammations)을 가진 환자의 데이터에서 특징을 추출하여 진료의 방향을 결정하는 학습을 진행한다. 특징이 추출된 데이터는 각 속성에 따른 인스턴스(instances)를 클러스터링(Clustering)을 적용한 나이브 베이안(Naive Bayesian)을 이용하여, 환자의 특징에 맞는 병명을 판단하기 위한 학습을 한다. 나이브 베이지안 학습은 각 속성별로 결합하는 확률이 서로 독립적(independent)이라는 가정[2]을 하는 방법으로 대량의 데이터를 다루는 것에 적합하며, 효율적이고 신뢰도가 높은 분류기로서 이용되고 있다. 그러나 의료데이터마이닝은 여타 데이터마이닝 분야에 비해서 까다로운 절차가 요구되는 분야로서, 그 수행과정이 보다 신중하게 진행되어야 한다. 따라서 높은 신뢰도를 확보하기 위해서 클러스터링 과정을 선행하여 클러스터링 클래스(Clustering-class)를 추가한다. 이로 인해서 병명에 대한 적합율의 신뢰도가 상승되었으며, 대량의 데이터에서는 수행시간이 단축되는 결과를 가져왔다.

2. 본 론

2.1 의료데이터마이닝

데이터마이닝은 주어진 데이터 속에서 규칙적인 패턴을 발견하고 의사결정에서 최선의 선택을 위해 정보를 활용하는 것이다. 따라서 대량의 데이터가 존재하는 비즈니스, 공학, 과학 그리고 의학 등 다양한 분야에서 효과적으로 활용되어 왔으며, 그 성과 또한 눈부시다.

견고한 예측모델의 생성과 신뢰성 높은 예측을 통하여 의료 실무자들에게 병의 예후와 진단에 도움을 주는 의료분야의 데이터마이닝은[3] 높은 신뢰도의 요구와 데이터의 특수성에 기인하여 활용도가 다소 미미하다. 그러나 각종 검사결과를 기반으로 자동적으로 진단명을 예측한다면 의료진들은 보다 쉽고, 체계적으로 환자에 대한 정확한 진단을 할 수 있다. 따라서 진단분야에서 데이터마이닝을 활용하기 위해서는 데이터의 특성을 고려한 보다 확실한 적합률이 필요하다.

2.2 나이브 베이지안 학습

나이브 베이지안 학습은 클래스의 사전 지식과 데이터로부터 획득한 새로운 증거를 결합시키는 통계 원리인 베이스 정리(Bayes theorem)에서 조건부 확률을 계산하기 위한 베이지안 분류 방법이다.[4] 클래스 레이블 y가 주어졌을 때, 나이브 베이지안 분류기는 속성들이 조건부로 독립적이라고 가정하여 조건부 확률로 계산하며, 조건부 독립성 가정은 다음과 같이 정의한다.

$$P(X|Y=y)=\prod_{i=1}^{d}P(X_i|Y=y) \qquad (1)$$

식(1)에서 속성 집합 $X=\{X_1,X_2,\ldots,X_d\}$는 d개의 속성으로 구성된다.[5] 그러나 이는 특정 분류 내의 클래스에서 인스턴스의 출현 확률이 같은 분류에 있는 다른 인스턴스의 출현 확률과 관련이 없다는 것이다. 즉 클래스별 인스턴스의 독립성이 유지된다는 의미이다. 이러한 가정은 실제 데이터 적용에서 근원적인 결함을 가지고 있으나 효율적인 분류방법으로 여러 분야에서 사용되고 있다. 하지만 의료데이터마이닝의 경우 분류의 신뢰도가 매우 중요함으로 나이브 베이지안 학습보다 베이지안 네트워크(Bayesian Networks) 학습[6]이 더 이용되고 있다.

2.3 제안한 클러스터링 클래스를 적용한 학습

클러스터링은 데이터를 특징이 정의되지 않은 집합으로 그룹화 하는 것으로 관심 있는 그룹의 개수나 구조를 고려하지 않고 이루어지는 분석 기법이다.[7]

본 논문에서는 나이브 베이지안 학습 이전에 전처리 과정으로 데이터를 클러스터링하여 클러스터링 클래스 C를 추가한다. 그 후 나이브 베이지안 분류기는 시험 항목을 분류하기 위해 각 클래스 Y에 대한 사후확률을 계산 한다.

$$P(Y|X)=\frac{P(Y)P(C|Y)\Pi_{i=1}^{d}P(X_i|Y)}{P(X)} \qquad (2)$$

식(2)는 $P(X)$가 모든 Y에 고정되어 있기 때문에 분모를 최대화하는 클래스 $P(Y)P(C|Y)\Pi_{i=1}^{d}P(X_i|Y)$를 선택하면 되고, 이는 클러스터링 클래스의 Y에 대한 사후확률을 포함한다.

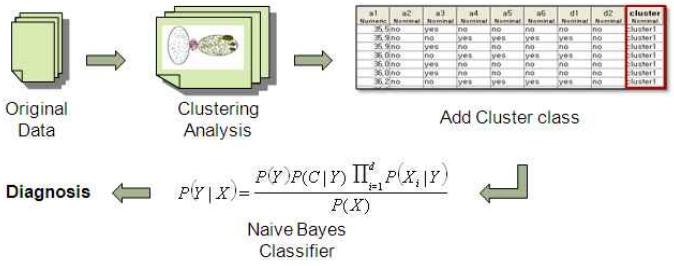

그림 2. 클러스터링 클래스를 적용한 학습 과정

위 그림은 오리지널 데이터를 특성별로 클러스터링 한 후 나이브 베이지안 방법을 통해 학습하는 과정을 나타내는 것이다. 본 논문은 위의 과정을 바탕으로 실험을 진행한다.

3. 실 험

클러스터링 클래스를 적용한 나이브 베이지안 학습 방법은 급성염증을 갖은 환자들의 데이터에 적용하였다. 데이터는 'diagnosis.data'[8]로 총 120개의 인스턴스에서 속성은 체온, 구토 증상, 요추 통증, 소변 배뇨, 방뇨 통증, 요도 팽창으로 총 6개를 추출하였으며, 병명의 판단은 급성 방광염과 결석으로 인한 신장염으로 제한하였다.

[표1] 클러스터링 클래스를 포함한 데이터

NO.	A1	A2	A3	A4	A5	A6	방광염	신장염	C_class
					:				
57	37.9	no	no	yes	yes	no	yes	no	cluster1
58	37.9	no	yes	no	no	no	no	no	cluster1
59	37.9	no	no	yes	yes	yes	yes	no	cluster1
60	37.9	no	no	yes	no	no	yes	no	cluster1
61	38.0	no	yes	yes	no	yes	no	yes	cluster2
62	38.0	no	yes	yes	no	yes	no	yes	cluster2
63	38.1	no	yes	yes	no	yes	no	yes	cluster2
					:				

데이터 전처리 과정을 통해 클러스터링 클래스의 추가가 선행적으로 이루어 졌으며, 후에 나이브 베이지안 학습을 수행하였다. 학습 프로그램은 오픈 소스로서 Java 언어로 개발된 데이터마이닝 프로그램인 'weka'를 이용하였다.

[표2] 나이브 베이지안과 성능비교

	NB	C-Class NB
Correctly Classified Instances	95.8333 %	100 %
Incorrectly Classified Instances	4.1667 %	0 %
Time taken to build model	0.03 seconds	0.01 seconds

[표2] 베이지안 네트워크와 성능비교

	BN	C-Class NB
Correctly Classified Instances	100 %	100 %
Incorrectly Classified Instances	0 %	0 %
Time taken to build model	0.05 seconds	0.01 seconds

실험결과 추출된 5가지의 분류 속성을 바탕으로 진행된 클러스터링 과정을 통해 2개의 클러스터로 구분되었으며, 그 후 클러스터를 포함한 나이브 베이지안 학습에서 기존의 나이브 베이지안 학습에 비해 우수한 신뢰도를 보임을 확인 할 수 있었다. 더불어 수행시간도 나이브 베이지안에 비해서 단축한 결과를 확인할 수 있었다.

또한 의료데이터마이닝의 진단분야에서 가장 널리 사용되고 있는 베이지안 네트워크와 비교했을 때도 신뢰도 면에서 뒤쳐지지 않으며, 오히려 수행성능에서는 다소 앞선 결과를 보여주었다.

4. 결 론

본 논문에서는 의료데이터마이닝 분야에서 요구되는 높은 신뢰도에 부합하기 위해, 나이브 베이지안 학습시 선행적으로 클러스터링 클래스의 추가를 제안하였다. 실험결과 수행성능면에서도 기존의 나이브 베이지안의 학습보다 우수했으며, 수행시간도 단축할 수 있었다. 뿐만 아니라 대표적인 진단 학습방법인 베이지안 네트워크와의 비교에서도 뒤떨어지지 않는 성능을 보여주었다. 그러나 베이지안 네트워크의 경우 시각적 다이어그램으로 표현되므로 보다 쉽게 해석할 수 있으며, 노드 사이의 상호작용을 쉽게 파악할 수 있는 장점이 있다. 이러한 이유로 환자의 임상적 증상을 바탕으로 그 관계에 대한 연구가 보다 수월하다. 하지만 시간을 고려해야하는 즉각적이고 인텔리전스한 환자진단 연구에는 대량의 데이터를 다루기에 적합하며, 베이지안 네트워크에 수행 신뢰도가 뒤지지 않은 클러스터링 클래스를 추가한 나이브 베이지안 방법이 효과적으로 적용될 수 있다. 향후 과제로는 다양한 의료데이터 적용을 통해 객관성 입증이 필요하며, 더 효율적인 성능을 얻기 위해서 클러스터링에서 가장 적절한 클래스의 수 조절에 대한 연구가 진행되어야 한다.

참 고 문 헌

[1] 정용규, 김인철, "효율적인 의료데이터마이닝을 위한 특징 축소와 베이지안망 학습", 한국지능정보시스템학회 추계학술대회, pp.258-265, 2002

[2] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers", Proceeding20th Internal Conference New York : ACM Press, pp.616-623, 2003

[3] Agrawal. R, Srikant. R, "Minig Sequential Patterns", In Proceedings of the 11th International Conference on data Engineering, pp3-14, 1995

[4] Ian H.Witten, Frank Eibe, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, pp.84-88, 2000

[5] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, ELSEVIER, pp.223-231, 2006
[6] David Heckerman, Dan Geiger, David M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data", Machine Learning, vol.20, pp.197-243, 1995
[7] 김용신, 클러스터별 인공 신경망 구축을 통한 데이터 마이닝 모델의 성능 향상, 아주대 대학원 석사논문, 2003
[8] http://cml.ics.uci.edu/