# 조건부 확률에 기반한 범주형 자료의 거리 측정

# A distance metric of nominal attribute based on conditional probability

이재호, 우종하, 오경환
서강대학교 컴퓨터학과

Jaeho Lee, Jongha Woo and Kyunghwan Oh
School of Computer Science, Sogang University
E-mail : popopome@ailab.sogang.ac.kr

**ABSTRACT**

유사도 혹은 자료 간의 거리 개념은 많은 기계학습 알고리즘에서 사용되고 있는 중요한 측정개념이다. 하지만 입력되는 자료의 속성들 중 순서가 정의되지 않은 범주형 속성이 포함되어 있는 경우, 자료간의 유사도나 거리 측정에 어려움이 따른다. 비거리 기반의 알고리즘들의 경우 - C4.5, CART - 거리의 측정없이 작동할 수 있지만, 거리기반의 알고리즘들의 경우 범주형 속성의 거리 정보 결여로 효과적으로 적용될 수 없는 문제점을 갖고 있다.

본 논문에서는 이러한 범주형 자료들간 거리 측정을 자료 집합의 특성을 충분히 고려한 방법을 제안한다. 이를 위해 자료 집합의 선험적인 정보를 필요로한다. 이런 선험적 정보인 조건부 확률을 기반으로한 거리 측정방법을 제시하고 오류 피드백을 통해서 속성 간 거리 측정을 최적화 하려고 노력한다. 주어진 자료 집합에 대해 서로 다른 두 범주형 값이 목적 속성에 대해서 유사한 분포를 보인다면 이들 값들은 비교적 가까운 거리로 결정한다. 이렇게 결정된 거리를 기반으로 학습 단계를 진행하며 이때 발생한 오류들에 대해 피드백 작업을 진행한다. UCI Machine Learning Repository의 자료들을 이용한 실험 결과를 통해 제안한 거리 측정 방법의 우수한 성능을 확인하였다.

Key words : 범주형 자료, 거리 측정, 유사도, 오류 피드백

## 1. 서 론

많은 기계학습 알고리즘들이 자료 사이의 유사도나 거리를 이용하여 자료를 분류하거나 목적 속성을 예측한다. 따라서 자료간의 거리가 자료를 구성하는 개별 속성들의 특징들을 잘 대표할 수 있도록 측정되어야 한다. 이 경우 자료를 구성하고 있는 속성들은 크게 수치형 속성과 범주형 속성들로 구분할 수 있다. 수치형 속성 값의 경우 자료 간의 거리를 어렵지 않게 결정할 수 있다. 유클리드 거리 측정 방법을 비롯하여 다양한 측정방법들이 제안되어 왔다[3]. 하지만 범주형 속성의 경우, 수치형 속성과는 달리 자료간의 거리 측정에 어려움이 있다.

범주형 속성의 경우, 간단한 거리 측정 방법으로 오버랩(overlap)이 제안되었다[1]. 오버랩 측정방법에 따르면, 두 속성 값이 같은 심볼을 갖는 다면 거리는 0, 다른 심볼을 갖는다면 거리는 1로 결정된다. 따라서 범주형 속성 간의 거리는 항상 0, 1의 두 값으로만 결정된다.

하지만 범주형 속성 값 간의 거리가 0, 1 이상의 의미로 해석될 수 있다. 예를 들어 사과의 사진을 인식하려고 할 경우, 범주형 속성 중 색 정보가 사용된다면 녹색과 빨강 사이의 거리가 녹색과 흰색보다는 좀 더 가까운 거리로 결정될 수 있다. 이런 결정은 자연스러우며 선험적인 지식을 갖고 있는 자료 집합에 대해 범주형 값들 사이의 다른 거리가 존재한다고 생각할 수 있다.

이런 점을 착안하여 Stanfill과 Waltz는 조건부 확률을 이용하여 범주형 속성 간의 거리를 측정하기 위한 Value Difference Metric(VDM)을 제안하였다[2]. 이에 기반하여 VDM과 다른 수치형 속성의 거리 측정 방식을 혼합하여 범주형 속성과 수치형 속성들의 거리를 측정하는 방법도 제안되었다[3]. 또한 함수 근사 측정(function approximation) 방법을 이용하여 범주형 속성값 간의 최적의 거리를 찾는 방법도 제안되었다[4]. 하지만 이 방법의 경우 전역 탐색 방법으로 인해 많은 계산 시간을 필요로 한다.

본 논문에서는 조건부 확률기반의 거리측정 방법을 확장하고 오류 피드백 방법을 제안한다. 이를 통해 범주형 속성 간 거리를, 자료가 내포하고 있는 특징을 이용하여 측정하고자 한다. 2절에서는 제안한 거리 측정 방법을 설명하고, 3절에서는 이를 기반으로한 실험 결과를 보인다. 4절에서는 결론과 향후 연구 방향에 대해 제시한다.

## 2. 범주형 자료의 거리 측정

### 2.1 Overlap

범주형 자료의 거리 측정을 위한 간단한 방법으로 오버랩(Overlap) 방법을 이용한다. 거리 측정은 다음과 같이 정의한다.

$$D_{overlap}(x,y)=\begin{cases}1 & (x\neq y)\\ 0 & (x=y)\end{cases}$$

입력 자료 $x$와 $y$의 심볼이 같다면 둘 사이의 거리는 0, 다르면 1로 결정한다.

### 2.2 조건부 확률에 기반한 범주형 자료의 거리 측정

조건부확률에 기반한 범주형 자료의 거리 측정으로 VDM(Value Difference Metric)이 Stanfill과 Waltz에 의해서 제안되었다. VDM에서 범주형 자료 값 간의 거리는 다음과 같이 정의된다.

$$D_{vdm}(x,y)=\sum_{c=1}^{C}\left|\frac{N_{a,x,c}}{N_{a,x}}-\frac{N_{a,y,c}}{N_{a,y}}\right|^{q}$$

$$=\sum_{c=1}^{C}\left|P_{a,x,c}-P_{a,y,c}\right|^{q}$$

- $N_{a,x}$: 학습 집합 중 속성 $a$에 대해서 $x$값을 갖는 인스턴스들의 수
- $N_{a,x,c}$: 학습 집합 중 속성 $a$에 대해서 $x$값을 갖고, 목적속성으로 $c$값을 갖는 인스턴스들의 수
- $C$: 목적속성들의 수
- $q$: 상수 값으로 대개의 경우 1과 2값을 갖음
- $P_{a,x,c}$: 속성 $a$가 $x$값을 갖고 있는 경우, 목적 속성이 $c$값을 가질 확률

$$P_{a,x,c}=P(c\mid x_a)=\frac{N_{a,x,c}}{N_{a,x}}$$

두 속성 값 $x$, $y$간의 거리를 $D_{vdm}$으로 측정하는 경우, 만약 두 속성 값이 목적 속성에 대해 비슷한 분포를 보인다면 거리가 비교적 가깝게 결정될 것이다. 하지만 이 측정 방법에서는 속성 값 자체가 갖는 고유한 특성을 반영하지 못한다. 즉, 속성 값 $x$, $y$가 확률상 같은 분포를 보일 지라도 속성 값이 가지고 있는 고유의 특징으로 인해 두 값 간에는 어느 정도의 거리의 차이가 존재한다.

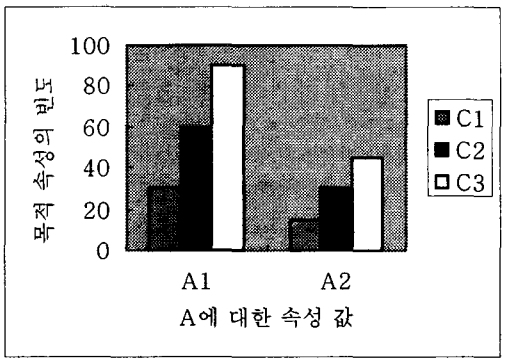


그림 1. 속성 A에 대한 목적속성 분포표

그림 1에서 보는 것과 같이 A1과 A2의 확률 분포는 서로 같지만, A1과 A2 값 각각이 목적 속성의 기여하는 정도가 다르다. 따라서 이런 경우, 속성 값 자체가 가지는 특징을 반영하도록 거리 측정 방법이 조정되어야 한다. 이를 위해 각각의 속성 값들이 목적속성에 얼마나 기여하는 정도를 속성 값 고유의 특징으로 사용한다.

또한 속성 값들의 분포를 맵핑 시킴으로 속성값 사이의 차이를 명확하게 할 수 있다. 이를 위해 속성 값이 가지는 확률값들을 로그 함수를 이용하여 맵핑시킨다. 따라서 위의 $D_{vdm}$ 식을 다음과 같이 확장하였다.

$$D_{cp}(x,y)=\sum_{c=1}^{C}\left|-log_2(P_{a,x,c})+log_2(P_{a,y,c})\right|^{q}$$

$$P_{a,x,c}=\frac{N_{a,x}}{N_T}\times\frac{N_{a,x,c}}{N_{a,x}}=\frac{N_{a,x,c}}{N_T}$$

- $N_T$: 총 인스턴스 수

$D_{cp}$와 $D_{vdm}$의 거리 값들은 각각의 속성별로 계산 되어지고 값들을 행렬에 저장된다. 결국 개별 속성마다 거리 값을 갖는 하나의 행렬을 갖게된다.

### 2.3 오류 피드백

$D_{cp}$의 거리 측정이 확률적 분포를 기반으로 측정된 값이므로 오차의 폭을 가지고 있다. 이 오차의 폭을 줄이기 위해 오류에 대한 피드백 작업을 진행한다. 이것은 전체 알고리즘의 학습단계에서 $D_{cp}$를 학습하기 위한 별도의 과정을 요구한다. 본 논문에서는 k-NN(k Nearest Neighbor) 알고리즘을 이용하여 $D_{cp}$를 학습, 검증하였다. k-NN을 적용한 결과 발생된 오류들을 기반으로 각 속성값 간의 거리를 갱신한다.

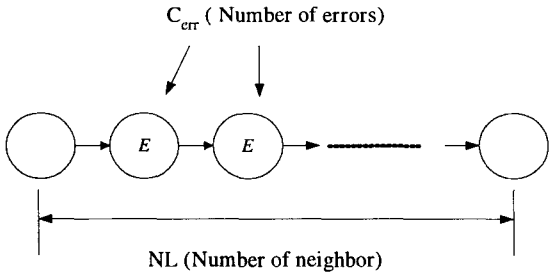


그림 2. 검증 인스턴스를 오류로 예측한 이웃들

그림 2는 검증 인스턴스(Test Instance)를 잘못 예측한 경우, 오류 피드백을 위해서 검증 인스턴스의 최근접 이웃들(Nearest Neighbor Instance)을 나타내고 있다. 이 최근접 이웃들을 이용하여 다음과 같이 정의된 식을 따라서 $x$와 $y$간의 거리를 갱신한다.

$$D_{cp}^{new}(x,y) = D_{cp}^{old}(x,y) + p \times \frac{C_{err}(NL)}{length(NL)}$$

- $NL$: 검증 인스턴스가 잘못 예측된 경우, 예측을 위해서 구성한 최근접 이웃들로 이루어진 목록
- $x$: 검증 인스턴스의 속성 $a$의 값
- $y$: $NL$에 있는 인스턴스들 중 검증 인스턴스와 다른 목적속성을 갖는 인스턴스의 속성 $a$의 값
- $C_{err}(NL)$ : $NL$중에서 오류로 결정된 인스턴스들의 수
- $p$: 매개변수 값

기본적으로 검증 인스턴스를 잘못 예측한 경우, 이 잘못된 예측이 많은 오류들을 바탕으로 이루어졌다면 이에 비례하여 더욱 큰 벌점을 각각의 속성 값들 사이의 거리에 부과하도록 한다. 검증 단계에서 매 반복마다 k-NN 알고리즘을 이용하여 오류들을 판별하고 이 오류들을 이용하여 피드백 작업을 진행한다. 오류 피드백은 최종적으로 오류 발생률이 증가하는 지점에서 중지한다. 그림 3은 학습 단계에서 $D_{cp}$를 결정하기 위한 전체 과정을 기술하고 있다.

1. Training step
   - A. Divide training data into two groups. One group is used to train and about 2/3. Other group is for test and about 1/3.
   - B. With this training group, calculate distances between values which are nominal.
   - C. Finally, each attribute has a matrix for distance between nominal values.
   - D. Test & Error feedback step
     - i. Get one instance from test group and test it with k-Nearest Neighbor.
     - ii. If the prediction is right, then skip to next instance and do again from step i. If there is no instance to test, exit training step.
     - iii. To get error feedback parameter value, find nearest neighbor list which is used to predict target attribute.
     - iv. Count errorneous instances and update distance of nominal attribute values with penalty parameter. Go to step i.
   - E. If error ratio which is determined through step D decreases, then there is more possibility to reduce error. So go to step D. If not, stop feed back and exit.

그림 3. $D_{cp}$ 계산 알고리즘

## 3. 실험결과

본 논문에서 제안한 조건부 확률 기반 접근의 유용함을 입증하기위해 다양한 자료 집합을 검증 자료로 이용하였다. 우선 각각의 자료 집합에 대해서 모든 결측값(Missing Value)을 갖고 있는 인스턴스들은 실험에서 제외되었다. 더불어 수치형 속성들도 자료 집합에서 제거되었다. 이렇게 함으로 결측값과 수치형 속성 값으로 인한 영향을 배제하였다. 사용한 자료 집합은 UCI Machine Learning Repository[5]에 있는 자료 중 범주형 속성값을 갖고 있는

자료들로 표 1과 같다.

| | Data set | Total Attributes | Nominal Attributes |
|---|---|---|---|
| Splice | 3190 | 62 | 61 |
| Grub & Damage | 155 | 9 | 7 |
| Balance scale | 625 | 5 | 5 |
| Soybean | 562 | 36 | 36 |
| Breast Cancer | 277 | 10 | 10 |

표 1. 데이터 집합

각각의 데이터 집합에 대해 세가지 알고리즘을 이용하여 실험하였다. 기본적인 k-NN 알고리즘 - 오버랩(overlap) 기반 알고리즘 -과 VDM을 기반으로 한 k-NN 알고리즘, 마지막으로 본 논문에서 제안한 조건부 확률을 기반으로 한 k-NN 알고리즘이다. 모든 데이터들에 대해서 10 fold cross-validation 사용하여 테스트 하였다.

본 논문에서 제안한 알고리즘을 적용하기 위해 학습 단계에서 학습 자료를 학습 그룹과 검증 그룹으로 분할 하였다. 학습 그룹은 본래 학습 자료의 2/3를, 검증 그룹은 1/3을 갖도록 하였다. 이 학습 단계에서 모든 자료들이 $D_{cp}$ 측정을 위해서 한번씩 검증되도록 3번 반복하였다. 모든 알고리즘은 각 집합에 대해 10번씩 반복 실행하였고 표2는 이들 알고리즘의 예측 성공률 결과를 보여준다.

| | Overlap k-NN | VDM k-NN | CP k-NN |
|---|---|---|---|
| Splice | 79.69 | 89.04 | **91.70** |
| Grub & Damage | 43.79 | 46.06 | **46.95** |
| Balance scale | **84.51** | 83.04 | 84.24 |
| Soybean | 89.02 | 87.93 | **89.71** |
| Breast Cancer | **75.41** | 73.94 | **75.41** |

표2. 예측 성공률(단위: %)

위의 실험결과로부터 5개의 데이터 집합들 중 조건부 확률 기반으로 한 CP k-NN가 전반적으로 좋은 예측률을 보이고 있다. 다른 알고리즘들과는 달리 안정적인 예측률을 보여준다. Balance scale과 Breast cancer 데이터 집합의 경우 Overlap k-NN도 좋은 성능을 보이고 있는데, 이 두 데이터 집합의 경우 대다수의 속성값들이 수치형 자료 집합을 일정 구간으로 잘라 범주형 값으로 변환된 특징을 갖고 있었다. 나머지 세 집합의 경우, 본논문에서 제안한 조건부 확률 기반의 접근이 좋은 성능을 보여주고 있다.

## 4. 결 론

본 논문에서는 선험적인 확률 분포 자료를 기반으로 범주형 자료간 거리 측정의 새로운 접근 방법을 제시하였다. 이를 기반으로 실험을 수행한 결과 기존 방법보다 우월한 성능을 보임을 입증하였다. 제안한 조건부 확률 기반의 범주형 속성 값 간의 거리 측정은 자료 집합의 선험적인 정보에 의존한다. 따라서 이 방법으로 결정된 범주형 자료 간 거리는 다른 자료 집합에 대해 유용하지 않다. 하지만 한번 특정 자료 집합에 대해 결정된 거리 값들은 거리 기반의 기계 학습 알고리즘들의 성능 향상에 매우 유용하게 사용될 수 있다.

그러나 다양한 자료 집합들을 통한 성능 평가, 오류 피드백의 지역적 최적화 문제, 자료간 거리의 과대적합(over-fitting) 문제에 연구가 추가로 필요하다.

**감사의 글**: 본 연구는 과학기술부 주관 뇌신경 정보학 사업에 의해 지원되었음.

## 5. 참고문헌

[1] Ian H.Witten, Eibe Frank, "Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann Publishers, 2000, Page(s): 193~201, 238~246

[2] Stanfill, C., and D. Waltz, "Toward memory-based reasoning.", Communications of the ACM, Vol.29, December 1986. Page(s): 1213~1228

[3] D. Randall Wilson, Tony R. Martinez, "Improved Heterogeneous Distance Functions", Journal of Artifical Intelligence Research 6, 1997, Page(s):1~34

[4] Victor Cheng, C.H.Li and C.K.Li, "Intra-feature metric matrices for nominal data pattern classification", Proceedings of the 9th International Conference on Neural Information Processing(ICONIP'02), Vol 5, Page(s): 2587~2589

[5] Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.