kNN 알고리즘에서의 속성 가중치 자동계산 방법

이강일*, 이창환*
*동국대학교 정보통신공학과
e-mail : {leeki816, chlee}@dgu.edu

an Automatic Calculation Method of Feature Weights in k Nearest Neighbor Algorithms

Kang-Il Lee*, Chang-Hwan Lee*
*Dept of Information and Communication Engineering,
Dong-Guk University

요 약

기억기반학습의 일종인 최근접 이웃(k nearest neighbor) 알고리즘은 과거의 데이터들 중에서 새로운 개체와 유사한 데이터들을 이용해서 새로운 개체의 목적 값을 예측하는 것이다. 이 경우 속성의 가중치를 계산하는 방식은 kNN의 성능을 결정하는 중요한 요소가 된다. 본 논문에서는 기존의 다른 이론들과 달리 정보이론에서 사용되는 엔트로피 개념을 이용해서 속성의 가중치를 이론적이고, 효과적으로 계산하는 새로운 방법을 제시하고자한다. 제안된 방법은 각 속성이 목적속성에 제공하는 정보의 양에 따라 가중치를 자동으로 계산하여 kNN의 성능을 향상시킨다. 마지막으로 이러한 방식의 성능을 다수의 실험을 통해 비교하였다.

1. 서론

기계학습 분야에는 여러 가지 예측/분류 알고리즘이 있으나 최근접 이웃 알고리즘(k nearest neighbor algorithm, 이하 kNN 알고리즘)은 그 중에서 적은 비용으로 비교적 높은 정확도를 제공하는 알고리즘이다. 최근접 이웃 알고리즘은 이미 알려진 개체들을 훈련 집합(training set)의 형태로 메모리에 기억한 다음 새로운 집합(test set)이 입력되었을 때, 저장된 개체 중에서 유사한 것들을 선택하여 그 개체들이 가진 목적 값에 따라 새로운 개체의 목적 값을 예측하는 방식의 분류 알고리즘이다.

이러한 kNN 알고리즘의 적용에 있어서 각 속성에 대하여 정확한 가중치 부여는 kNN 알고리즘의 성능에 많은 영향을 미친다. 즉 모든 속성을 같은 비중으로 판단하게 되면 kNN 알고리즘은 높은 성능을 제공할 수 없다. 예를 들어서, 학생 데이터베이스에서 특정 학생이 어떠한 대학에 대한 합격 가능 여부를 알아볼 때, 그 학생의 수능성적 속성은 주소 속성에 비해 더 높은 가중치를 부여 받아야한다. 반대로 그 학생이 집에서 특정 목적지에 도착하는 교통편을 알기 위해서는 반대로 주소 속성이 더 큰 가중치를 부여 받아야한다. 이처럼 데이터베이스에서 특정한 목적속성(target attribute)의 값을 예측/분류하고자 할 때, 목적속성을 제외한 다른 속성의 가중치는 목적속성에 따라서 그 값을 달리하게 된다. 이와 같은 이유로 kNN 알고리즘에서 각 속성에 대한 정확한 가중치 부여는 알고리즘 성능에 있어서 많은 영향을 미치게 되는 부분이 된다.

본 논문에서는 정보이론(information theory) 중에서 Hellinger 변량을 이용하여 각 속성의 가중치를 자동으로 계산하는 새로이 개선된 알고리즘을 소개하고자 한다.

2. 관련 연구

kNN 알고리즘은 최초로 Cover와 Hart[1]에 의하여 제안되었다. 이 후 Smith와 Medin[2]등에 의하여 타당성을 인정받았고 이후 Aha, Kibler and Albert[3]에 의하여 몇 개의 개체중심 학습(instance-based learning, IBL) 알고리즘이 개발되었고, 이들은 각각 IB1, IB2, IB3 로 알려져 있다.

Zhang[4]은 개체간의 유사도와 개체내의 유사도를 정의하여 계산된 대표적인 개체를 개념설명의 형태로 처리하는 알고리즘을 발표하였다. Salzberg가 제안한 EACH[5]에서는 학습 데이터의 분류 이후에 정확한 분류에 대해서는 속성의 가중치를 일정한 양만큼 향상시키고 반대의 경우 일정한 양만큼 감소하는 방법을 사용하였다.

Creecy[6]등은 조건 확률을 사용하는 가중치 계산법을 제안했는데, 이는 전체 속성을 이진 속성으로 전환하고 각 속성에 대해 PCF(per-category feature importance)를 계산하여 가중치를 구했다.

또한, Cost와 Salzberg이 제시한 PEBLS[7] 에서는 MVDM(Modified Value Difference Metric) 방법을 제시하였는데 속성 값의 분포가 한 곳으로 치우칠수록 높은 가중치를 가지도록 설계되어있다.

본 논문은 이러한 kNN 알고리즘에서 속성의 가중치 계산을 수행함에 있어, 위의 방법들과 달리 정보이론의 엔트로피 함수를 이용하여 각 속성이 목적값에 미치는 영향을 정량적으로 분석하며, 그에 따라 각 속성의 가중치를 계산한다. 따라서 각 속성의 가중치에 대하여 이론적인 배경을 제공하며 좀 더 정확한 가중치 값을 제공할 수 있다.

3 kNN 알고리즘의 내용

kNN 알고리즘의 수행을 위해서는 속성의 가중치 계산, 속성 값 간의 유사도 계산 등의 계산법을 필요로 한다. 이 장에서는 본 논문에서 사용하는 kNN 알고리즘의 기본적인 내용을 설명한다.

우선 X와 Y를 각각 k 개의 속성으로 구성된 데이터라고 할 때, x_i와 y_i를 각각 X와 Y의 i번째 속성 값이라고 하자. 또한 T를 목적 속성이라고 가정하고 $\Delta_T(X, Y)$를 T에 대한 X 와 Y의 유사도라고 할 때, $\Delta_T(X, Y)$는 다음과 같이 정의된다.

$$\Delta_T(X, Y) = \sum_{i=1}^{k} \omega_T(i) \cdot S_T(x_i, y_i)$$

여기서 $S_T(x_i, y_i)$는 속성 값 x_i, y_i 간의 유사도이며 $w_T(i)$는 i번째 속성의 가중치이다.

위의 식에서 볼 수 있듯이 kNN 알고리즘에서 유사도의 계산은 다음의 두 가지 단계로 구분된다.

1. 속성 값의 유사도($S_T(x_i, y_i)$)의 계산
2. 각 속성의 가중치($w_T(i)$) 계산

위의 식들을 기반으로 새로운 데이터에 대해 과거의 데이터에서 가장 유사도가 높은 데이터들을 선택해 그들의 목적 값을 이용하여 분류를 수행한다.

3.1 속성에 따른 유사도의 계산

속성 내에서 속성의 값들 간의 유사도를 계산하는 방법은 속성의 타입에 따라서 달라지며 크게 세 가지 타입중의 하나로 구성되어 있다: 이진 속성, 연속 속성, 카테고리 속성. 본 논문의 주제는 속성의 가중치 자동 계산에 대한 내용이지만 kNN 알고리즘을 실제 구현하기 위해서는 데이터의 속성에 따른 유사도를 정의하여야 하며 이 장에서는 구체적인 방법을 설명한다.

먼저 이진속성 경우, 임의 속성 A에서의 두 값을 x_i와 y_i라하고 이 값들 간의 유사도를 $S_T(x_i, y_i)$라고 할 때 이진 속성의 경우는 두 개체의 값이 같은 경우 1 그렇지 않은 경우 0 을 부여한다.

특정한 속성 A가 연속 변수인 경우는 두 변수를 n-차원의 유클리드 공간(Euclidean space, E^m)에서의 두 점으로 간주하는 방법이 가장 많이 사용하지만, 이 경우 유사도가 값의 범위에 따라 큰 변동을 보이므로 유사도의 값을 정규화 시켜 사용한다.

카테고리 속성에서의 유사도 계산을 위해서 본 논문에서는 데이터 값들 간의 현재 가장 많이 알려진 Stanfill과 Walz의 VDM(Value Difference Metric)의 통계적인 방법을 이용하여 카테고리형 데이터의 모든 값의 조합에 대하여 유사도를 정의하였다[8].

3.2 속성 가중치의 계산방법

이 장에서는 본 논문의 주된 주제인 kNN 알고리즘의 속성가중치 계산방법에 대하여 설명하고자 한다.

본 논문에서 제안된 속성의 가중치를 계산하는 방법은 다음의 세 가설에 바탕을 두고 있다: (1) 속성의 특정한 값이 정해지면 이는 목적 속성에 정보를 제공한다. (2) 제공되는 정보의 양은 엔트로피 함수에 의하여 정의될 수 있다. (3) 속성이 제공하는 정보의 양이 많을수록 엔트로피 함수의 값은 커진다. 이분 속성, 카테고리 속성과 같은 이산형(discrete)의 속성에서 속성에 대한 가중치의 계산을 위해서는 먼저 속성의 각 이산형 값에 대한 정보량을 계산한 후 평균값을 해당 속성의 가중치로 사용한다. 연속 속성의 경우에는 데이터를 몇 개의 범위로 분할하여 주는 이산화과정(discretization)을 수행

후 그 결과를 가지고 위의 방법을 적용하는 것이다.

이제 가장 중요하게 남은 것은 '속성의 특정 값이 목적 속성에 제공하는 정보의 양을 어떤 방법으로 측정할 것인가' 이다. 이 글에서는 정보의 양을 측정하는 엔트로피 함수로서 Hellinger 변량(divergence)을 사용하는데 이는 Beran[15]에 의해 제안된 후 여러 분야에서 사용되고 있다.

목적 속성을 T라고 하고, '특정 속성 A의 값이 a가 됨'을 'A=a'로 표현한다. t_i를 T값 중의 하나로 가정하고, $p(t_i)$와 $p(t_i|a)$를 목적 속성 T의 사전 확률 및 사후 확률로 가정한다. A=a가 T에게 제공하는 정보의 양을 $H(T|A=a)$로 표시할 때 그 변량은 다음과 같이 정의된다.

$$H(T|A=a)=\left[\sum_i\left(\sqrt{p(t_i)}-\sqrt{p(t_i|A=a)}\right)^2\right]^{\frac{1}{2}}$$

이 변량은 목적 속성 T의 사전 확률분포와 사후 확률분포간의 차이를 측정한 함수이다. 이 변량의 특성에 대하여 조사하여 보면 우선 모든 경우의 $p(t_i)$와 $p(t_i|a)$ 값에 대하여 그 값이 정의가능(definable)하고 연속(continuous)이다.

또한 위의 Hellinger 변량은 사전 확률분포와 사후 확률분포가 일치할 때만 값이 0이 되며 나머지 경우는 항상 0과 1사이의 값을 가진다. 특정 속성 A의 가중치 계산을 위해 A가 가지는 모든 값에 대하여 위에 정의된 Hellinger 변량 값을 구하고 그 합을 A의 가중치로 사용할 수 있을 것이다. 하지만 이 경우 속성이 가지는 값의 개수가 증가하게 되면 변량의 값도 증가하게 된다. 본 연구에서는 위 변량에 각 속성 값의 발생확률 $P(a)$를 곱해 해결하였다.

$$H(T|A)=\sum_a p(a)\cdot H(T|A=a)$$

따라서 $H(T|A)$의 값은 속성의 값의 개수에 영향을 받지 않게 된다. 끝으로 위에서 정의된 가중치의 범위를 0과 1사이로 제한하기 위하여 모든 속성의 가중치 값의 합에 대한 비율로써 표현하였다. 최종적인 가중치의 값의 식은 다음과 같이 표현된다.

$$\omega_T(A)=\frac{H(T|A)}{\sum_A H(T|A)}=\frac{\sum_a p(a)H(T|A=a)}{\sum_A H(T|A)}$$

kNN 알고리즘을 이용한 분류학습에 있어서 흔하게 발생하는 것 중의 하나는 데이터의 누락치 문제이다. 본 연구에서는 누락치가 가지고 있는 정보의 양을 계산한 다음 이를 속성내의 다른 값들에게 빈도수에 비례하여 분배하는 방식을 사용하였다.

제안된 알고리즘은 분류하고자하는 개체를 훈련 집합 내의 각 개체에 대하여 유사도를 계산한 후 가장 유사도가 높은 몇 개의 개체를 선택하여야한다. 오직 한 개의 가장 유사한 개체만을 선택하면 에러나 노이즈(noise)에 심하게 영향을 받을 수 있으므로 두 개 이상의 개체를 선택하여 종합적으로 비교하는 것이 바람직하다.

본 연구에서는 각 개체의 유사도를 각 개체의 가중치로 사용해 유사한 개체가 더욱 높은 가중치를 가질 수 있게 하였다.

4. 실험 결과

본 연구에서 제안된 가중치 계산의 방법은 C언어로 구현되었다. 또한 성능을 평가하기 위해 동일 데이터에서 가중치를 고려하지 않은 상태(모든 속성의 가중치를 1.0으로 동일하게 설정)에서의 분류학습과 그 성능을 비교하였다.

성능 비교를 위하여 실험에 사용한 데이터는 University of California Irvine의 데이터 집합소[9]에서 다음 5개의 데이터 집합을 선택하였다: Breast Cancer, Echocardiogram, Liver Disorders, Pima Diabetes, Voting. 알고리즘의 테스트는 훈련 집합/테스트 집합의 방법을 이용한다. 첫째 전체 데이터의 70%를 임의로 선택하여 훈련 집합으로 하고 나머지를 테스트 집합으로 한다. 둘째로 훈련 집합에 대하여 위에서 설명한 두 개의 알고리즘들을 수행한 후 테스트 집합의 각 개체에 대하여 목적 속성의 값을 예측한다. 셋째로 테스트 집합의 모든 개체에 대하여 예측결과를 실제 결과와 비교 그 정확도를 계산한다. 편차를 줄이기 위해 훈련 집합과 테스트 집합의 분할을 10회 반복 수행하였다. 또한 데이터가 연속 속성을 포함할 때에는 연속속성 값을 구간 값으로 변환하는 이산화를 필요로 한다(속성 값의 유사도 계산에서는 연속속성의 값을 그대로 사용함). 연속속성의 이산화 방법은 여러 방법들이 제시되어 있지만 본 연구의 실험비교에는 서로 같은 이산화 방법을 사용하는 한 수행 결과에 영향을 주지 않으므로 편의상 각 연속속성의 값은 동일빈도 방식으로 5등분하여 이산화 하였다.

표 1은 Breast cancer 데이터에 대해 계산된 속성의 가중치이며 표 2는 Echocardiogram 데이터의

각 속성에 대한 가중치를 보여주고 있다. 표 3은 각 데이터에 대하여 속성의 가중치를 고려한 알고리즘과 그렇지 않는 알고리즘의 정확도를 비교하고 있다. 각 실험의 정확도 값은 데이터를 7:3 비율로 분할하는 경우를 10회 반복하여 수행한 평균값을 의미한다. 보는바와 같이 속성의 가중치를 자동 계산하는 알고리즘은 그렇지 않은 방법에 비해 높은 정확도를 가진다는 것을 알 수 있다. 이들 데이터 중 Breast cancer, Echocardiogram, Voting 데이터는 누락치를 포함하며 따라서 이러한 처리 결과는 본 연구의 계산방법이 누락치를 잘 처리하고 있음을 보여준다.

속성	가중치
Clump thickness	0.0988
Uniformity of cell size	0.1444
Uniformity of cell shape	0.1443
Marginal adhesion	0.0986
Single epithelial cell size	0.1127
Bare nuclei	0.1305
Bland chromatin	0.1207
Normal nucleoli	0.1040
Mitoses	0.0460

표 1 : Breast cancer 데이터의 속성가중치

속성	가중치
Age-at-heart-attack	0.1410
Pericardial-effusion	0.0272
Fractional-shortening	0.0891
Epss	0.1868
Lvdd	0.1634
Wall-motion-score	0.1869
Wall-motion-index	0.2056

표 2 : Echocardiogram 데이터의 속성 가중치

데이터	가중치 사용	가중치 미사용
Breast cancer	96.6	94.3
Echocardiogram	90.7	81.6
Liver Disorders	69.4	62.5
Pima Diabetes	74.6	70.3
Voting	96.2	89.8

표 3 : 가중치 계산방법의 정확도 비교(단위 : %)

5. 결 론

본 논문에서는 kNN 알고리즘의 성능에 많은 영향을 미치는 속성의 가중치 계산방법으로 정보이론을 바탕으로 한 엔트로피 함수의 일종인 Hellinger 변량을 새로운 가중치 계산 방법으로 채택하여 개발하게 되었다. 제안된 방법의 효율성을 실효성 여부를 검정하기 위해 우리는 5가지의 데이터베이스를 선택하여 실제 적용하여 보았고 그 결과를 가중치를 고려하지 않는 방법과 비교를 통해 그 타당성을 증명하였다. 제안된 가중치 계산 방법은 모든 경우에 대하여 더 나은 성능을 보였다.

본 논문에서 개발된 알고리즘 특징인 속성 가중치의 자동부여 방법은 kNN 알고리즘의 응용범위를 대폭 확장시킨다. 이는 다른 기계학습 방법에 비해 적은 비용으로 비교적 높은 정확성을 보이는 kNN 알고리즘의 중요성에 비춰볼 때, 기계학습 분야에 대한 발전을 꾀할 수 있을 것으로 기대한다.

참고문헌

[1] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, Vol. 13

[2] E. E. Smith and D. L. Medin, "Categories and Concepts" Cambridge, MA: Harvard University Press

[3] D. Aha, D. Kibler and M. Albert, "Instance-based Learning Algorithms," Machine Learning

[4] J. Zhang, ``Selecting typical instances in instance-based learning,'' Proceedings of the Ninth Int. Machine Learning Conference, Aberdeen, Scotland: Morgan Kaufmann

[5] S. Salzberg, "A Nearest Hyperrectangle Learning Method," Machine Learning

[6] R. Creecy, B. Masand, S. Smith, and D. Waltz "Trading MIPS and Memory for Knowledge Engineering," Communications of the ACM

[7] S. Cost and S. Salzberg, "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features," Machine Learning

[8] C. Stanfill and D. Waltz, ``Toward Memory-based Reasoning,'' Communications of the ACM

[9] P. Murphy and D. Aha, UCI Repository of Machine Learning Databases, Irvine, CA : University of California Irvine, Department of Information and Computer Science