k-최근접 이웃 정보를 활용한 베이지안 추론 분류

노영균, 김기응, 이태훈*, 윤성로*, Daniel D. Lee**
한국과학기술원, 서울대학교*, 펜실베니아 대학교**

요 약

본 리뷰 논문에서는 많은 데이터 환경에서 얻어진 k-최근접 이웃들(k-nearest neighbors)의 이론적 성질로부터 어떻게 분류를 위한 알고리즘을 만들어낼 것인가에 대한 여러 가지 방법들을 설명한다. 많은 데이터 환경에서의 최근접 이웃 데이터의 정보는 다양한 기계학습 문제를 푸는데 아주 좋은 이론적인 성질을 가지고 있다. 하지만, 이런 이론적인 특성들이 데이터가 많지 않은 환경에서는 전혀 나타나지 않을 뿐 아니라 오히려 다른 다양한 알고리즘들에 비해 성능이 많이 뒤쳐지는 결과를 보여주고 있다. 본 리뷰 논문에서는 많은 데이터 환경 하에서 k-최근접 이웃들의 정보가 어떤 이론적인 특성을 가지는지 설명하고, 특별히 이런 특성들을 가지고 k-최근접 이웃을 이용한 분류 문제를 어떻게 베이지안 추론(Baysian inference) 문제로 수식화 할 수 있는지 보인다. 마지막으로 현재의 빅데이터 환경에서 실용적으로 사용할 수 있는 알고리즘들을 소개한다.

Ⅰ. 서 론

k-최근접 이웃(k-nearest neighbor)을 사용한 분류 문제는 기계 학습과 패턴 인식 분야에서 가장 오래되고 많이 사용된 방법의 하나이다. 데이터 공간 상에서 가장 유사도가 높은 k개의 샘플을 골라 다수결을 통해 항목(class)를 결정하는 k-최근접 이웃 분류 방법은 알고리즘의 단순함 뿐만 아니라 뛰어난 일반화 능력으로 인해 데이터의 특성에 상관 없이 여러 측면에서 유용하게 이용되어 왔다.

하지만 많은 경우, 실제 문제를 푸는 과정에서는 일반적으로 k-최근접 이웃 분류법이 다른 많은 알고리즘들 보다 좋은 성능을 내지 못해 왔다. 따라서 k-최근접 방법은 초벌 분석(preliminary analysis)을 위한 도구로 사용되거나 다른 알고리즘이 상대적으로 얼마나 잘 하는지를 나타낼 때 쓰이는 최소 기준을 제공하는 베이스라인 분류기(baseline classifier)로 많이 사용되어 왔다.

반면, k-최근접 분류기가 이론적으로 얼마나 좋은 분류기가 될 수 있는지에 관한 연구는 이 방법을 데이터의 모분포(underlying probability density function)를 알 때에만 얻을 수 있는 베이즈 에러(Bayes error)와 연관시켜 베이즈 에러에 상당히 접근할 수 있음으로 보여준다[2],[3]. 단, 필요한 조건은 분류하고자 하는 데이터의 위치와 k개의 최근접 이웃의 위치들에서 모분포 값이 거의 같다고 할 수 있을 정도로 데이터의 밀집도가 큰 상황, 즉 데이터가 아주 많은 상황을 가정하고 있다. 이 경우에 최근접 이웃들을 이용한 분류는 단 한 개의 최근접 이웃들만 가지고 분류를 한다고 하더라고 베이즈 에러의 두 배를 넘지 않는 아주 작은 에러 기대값을 가지게 된다[2]. 또한 더 많은 개수의 최근접 이웃을 사용할수록 에러 기대값은 단조감소하여 결국 베이즈 에러에까지 접근한다는 것을 알 수 있다.

본 논문은 한걸음 더 나아가 최근접 이웃들까지의 거리들이 특정 확률 분포를 따른다는 데 주목한다. 이 확률 분포는 k번째의 최근접 거리를 반지름으로 하는 초구(hyper-shpere)의 체적(volume)을 확률변수(random variable)로 삼았을 때, 이 체적이 감마 분포(Gamma distribution)를 따르는 성질을 띈다[5]. 이 논문에서는 이러한 감마 분포를 통해 어떻게 k개의 최근접 이웃을 통한 단순한 다수결 방법이 최대 우도 추정법의 결과로 설명되는지 소개하고 여기에 더해서 베이지안 추론을 통해 더 좋은 전략을 고안해 낼 수 있을지를 고찰해 본다. 결과적으로 베이지안 추론법을 통해 우리는 주어진 최근접 이웃들의 정보가 주어졌을 때, 특정 항목(class)의 모분포 밀도함수가 다른 항목의 모분포 밀도함수보다 클 확률값을 닫힌 해(closed-form solution)로 얻을 수 있게 된다. 따라서, 분류 전략을 세우는데 있어서 이 확률값을 어떻게 활용하는지에 대한 논의가 이루어질 것이다.

마지막으로, 이러한 추론법을 통해 얻어지는 확률값을 보다 빨리 계산하기 위한 계산 방법을 소개한다. 주어진 k-최근접 이웃 정보를 가지고 항목(class) 사후 확률을 계산할 때, 일반적으로 적용되는 계산을 위해서 재귀법으로 프로그램을 구현 필요가 있다. 이 부분에 대해 효과적인 알고리즘의 구현을 위해

사용할 수 있는 기술을 잠시 소개한다.

본 리뷰 논문의 구성은 다음과 같이 구성 되어 있다. 두 번째 절에서는 많은 데이터 환경에서의 최근접 이웃의 정보가 어떤 특성을 지니는 지에 대한 이론적 리뷰를 제공한다. 세 번째 절에서는 이러한 특성을 이용해서 어떤 확률적 추론을 할 수 있는지 보여주고, 네 번째 절에서 이러한 확률 추론을 바탕으로 어떤 분류 전략을 세울 수 있을 것인지 설명한 다음, 다섯 번째 절에서는 간단한 실험들을 통한 알고리즘의 유용성을 제시한다. 여섯째로 어떻게 알고리즘을 다중 항목으로 확장시킬 지 설명하고 일곱번째 절에서 결과와 함께 끝맺는다.

Ⅱ. 많은 데이터 환경에서의 최근접 이웃 정보의 특성

Ⅱ-1. 베이즈 에러와 최근접 이웃 분류기

최근접 이웃 정보에 대한 이론적 연구는 최근점 이웃 정보를 모분포에 연결시켜 베이즈 에러와의 관계를 밝힌 T. Cover의 연구를 기반으로 시작되었다[2]. 이 연구는 데이터가 많아 데이터 공간을 충분히 메울 만한 밀도가 보장되는 가정 하의 결과였고, 결과적으로 알고자 하는 데이터 위치에서의 확률 밀도와 이 위치에서 가장 가까운 k개의 최근접 이웃들 위치에서의 확률 밀도 차이가 없게 되는 상황을 가정하였다.

즉, 모분포 $P(\mathbf{x}, y)$에서 생성된 N개의 데이터 $\mathcal{D} = \{\mathbf{x}_i, y\}_{i=1}^{N}$가 학습용 데이터로 주어진 상황을 생각했을 때, 분류를 하고자 하는 D 차원 상의 데이터 $\mathbf{x} \in \mathbb{R}^D$의 k개의 최근접 데이터 $\mathcal{D}_{NN} = \{\mathbf{x}_j, y_j\}_{j \in NN_k(i)}$들은 다음과 같은 사후 확률을 가진다. 간단히 두 개의 분류항목으로 가지고, 두 분류항목의 사전확률이 같아 $P(y=1) = P(y=2)$을 만족시킨다고 가정 했을 때, 데이터가 많은 상황에서, $\mathbf{x}_j$의 분류항목 y_j가 분류항목 1을 가질 확률과 분류항목 2를 가질 확률은 각각

$$P(y_j = 1|\mathbf{x}_j) \simeq \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=1) + p(\mathbf{x}|y=2)}$$
$$P(y_j = 2|\mathbf{x}_j) \simeq \frac{p(\mathbf{x}|y=2)}{p(\mathbf{x}|y=1) + p(\mathbf{x}|y=2)}$$

와 같이 근사 된다. 여기서, $\mathbf{x}_j \neq \mathbf{x}$ 임을 주목해야 한다.

T. Cover는 이 사후확률을 데이터마다 독립적으로 적용하여 데이터가 많을 때의 최근접 이웃 분류법의 에러의 기대값을 계산했다. 이 계산은 단순히 최근접 이웃의 인덱스가 j일 때, $\mathbf{x}$가 분류 항목과 $\mathbf{x}_j$의 분류 항목이 서로 다를 경우를 계산하는데, 이 확률이 바로 $\mathbf{x}$를 분류할 때의 최근접 이웃 분류기의 에러율이다.

에러의 확률이 $\mathbf{x}$가 분류항목 1일 확률과 $\mathbf{x}_j$가 분류항목 2일 확률, 그리고 $\mathbf{x}$가 분류항목 2일 확률과 $\mathbf{x}_j$가 분류항목 1일 확률의 곱이므로, 각 지점 $\mathbf{x}$에서의 에러 확률을 다음과 같이 계산하고,

$$P(y_j = 1|\mathbf{x}_j)P(y=2|\mathbf{x}) + P(y_j = 2|\mathbf{x}_j)P(y=1|\mathbf{x})$$
$$\approx \frac{2p(\mathbf{x}|y=1)p(\mathbf{x}|y=2)}{(p(\mathbf{x}|y=1) + p(\mathbf{x}|y=2)^2},$$

이를 $\mathbf{x}$의 밀도에 대해 기대값을 취하면, 최근접 이웃 분류법의 에러 기대값을 아래와 같이 구할 수 있다.

$$\epsilon = \int \frac{2p(\mathbf{x}|y=1)p(\mathbf{x}|y=2)}{(p(\mathbf{x}|y=1) + p(\mathbf{x}|y=2)} \mathrm{d}\mathbf{x}.$$

또한 이런 식으로 k-최근접 이웃 분류기의 다수결 전략이 만드는 에러를 만드는 경우의 수를 따져 에러의 기대값을 계산해 볼 수 있는데, T. Cover는 이 기대값이 최근접 이웃 분류기에 대해서는 언제나 베이즈 에러의 두 배를 넘지 않으며, k가 증가할수록 이 기대값은 단조감소하여 베이즈 에러까지 감소하다는 것을 보였다[3].

이 에러의 기대값은 아주 훌륭해서 분류기의 이론적 최소값인 베이즈 에러와 직접 연관이 있으며 (최악의 경우 베이즈 에러의 두 배), 두 분류 항목의 항목-조건부 모분포(class-conditional underlying density)가 많이 겹치지 않는 경우 베이즈 에러 자체가 낮아 최근접 이웃 분류법 만으로 좋은 성능을 낼 수 있다. 이것이 데이터가 많을 경우 최근접 이웃 분류기가 행동하는 방식으로 T. Cover가 제공한 이론적 설명이다.

Ⅱ-2. 최근접 이웃까지의 거리의 분포

최근의 연구에서는 최근접 이웃 정보를 정보이론 연구에 사용하는 시도들이 나타나고 있다. 이 연구들에서 공통적으로 사용하는 개념은 데이터가 모분포의 확률 밀도 함수로부터 무작위로(randomly) 생성 되었다면, 데이터 공간 안에서 최근접 이웃이 나타난 지점을 표면으로 하는 초구(hypersphere)의 부피가 지수분포 밀도 함수(Exponential density function)를 따른다는 것이다. 이는 확률이 확률 밀도 함수의 공간 적분에 비례하기 때문인데, 무작위성을 대표하는 포아쏭 프로세스(Poisson process)에서 첫번째 시그널이 나타나는 시점이 지수분포 밀도 함수를 따르는 것과 같은 현상이다.

최근접 이웃까지의 거리가 d 일 때, 반지름이 d 인 초구(hyper-sphere)의 체적(volume)을 γd^D 라고 할 수 있고 (단, $\gamma = \frac{\pi^{\frac{D}{2}}}{\Gamma(1+\frac{D}{2})}$), 데이터의 수가 N 이라고 할 때 새로운 확률

변수(random variable) $u = \gamma N d^D$ 을 정의할 수 있다. 데이터가 확률 밀도 λ 를 가지고 생성되었다고 하면, 이 새로운 체적에 대한 확률 변수는 다음과 같은 지수 분포 밀도 함수를 따른다:

$$p(u|\lambda) = \lambda \exp(-\lambda u)$$

데이터가 많은 경우, 확률 밀도가 일정한 구간에서는 최근접 이웃까지의 거리에 대한 정보로 위와 같은 density 정보를 사용할 수 있다. 이에 대한 단순한 확장으로 k번째 최근접 이웃까지의 거리는 다음과 같은 통계적 성질을 가진다. 즉, k번째 최근접 이웃까지의 거리가 d_k일 때, 새로운 확률 변수 $u_k = \gamma N d_k^D$ 가 가지는 확률 밀도 함수는 〈그림 1〉과 같이 차수가 k인 감마 분포 밀도 함수(Gamma density function), 혹은 k가 자연수인 걸 고려해 엘랑 분포 밀도 함수(Erlang density function)을 따른다고 할 수 있다.

$$p(u_k|\lambda) = \frac{\lambda^k}{\Gamma(k)} \exp(-\lambda u_k)(u_k)^{k-1}. \quad (1)$$

단, λ 는 최근접 이웃을 찾기 원하는 지점에서의 확률 밀도 함수값이다.

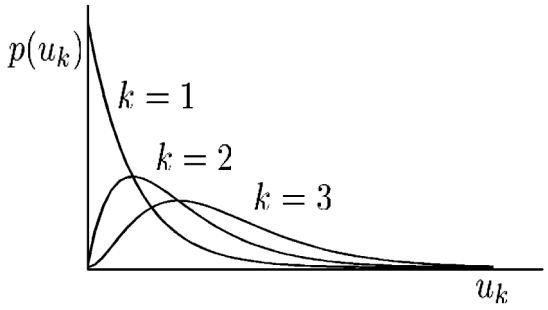


그림 1. k번째 최근접 이웃을 표면에 놓는 초구(hypersphere)의 체적에 대한 확률 밀도 함수

Ⅲ. 최근접 이웃 정보를 이용한 추정과 추론

Ⅲ-1. 최대 우도 추정

최근접 이웃까지의 거리가 모분포의 밀도값 λ 가 정해졌을 때 어떤 통계적 특성(stochastic property)을 지니는지 파악이 되었으면, 분류 문제를 푸는 데 있어서 이를 어떻게 응용할 것인지 생각해 볼 수 있다. 식 (1)의 형태로부터 자연스럽게 생각해 볼 수 있는 것은 데이터로부터 확률 밀도를 추정해 보는 최대 우도법을 적용시키는 문제이다. 즉, 모분포의 밀도를 파리미터(λ)로 놓고, 측정된 최근접 이웃들까지의 거리로부터 λ 를 측정하는 것이다. 우리는 각각의 분류항목이 가지는 최근접 이웃 정보들로부터 C 개의 분류항목 각각의 밀도값의 추정값 $\hat{\lambda}_1, \ldots, \hat{\lambda}_C$ 을 구하여 가장 큰 밀도값의 추정값을 가지는 분류항목을 택할 수 있다.

$$y = \arg\max_c \hat{\lambda}_c \quad (2)$$

이러한 접근법이 이전 k-최근접 이웃 분류에서의 다수결 방법과 다른 점은, 최근접 이웃 정보들을 가지고 보다 이론적이고 원리적인(principled) 분류 방법을 고안해 내었다는 데 있다. 특히나 최근접 이웃의 개수를 세는 것뿐만 아니라 거리를 이용하는 것, 그리고 각각의 최근접 이웃에서 얻은 정보가 어떻게 분류항목 결정에 기여하는 지 등의 분석과 이해를 제공한다는 면에서 의미있는 새로운 접근법이라고 할 수 있다.

정리 1 (비교방법과 다수결 방법의 동일성): 두 가지 분류 항목을 갖는 이진 분류 문제에서, 첫 번째 분류 항목의 k'번째 최근접 이웃까지의 거리와 두 번째 분류 항목의 k'번째 최근접 이웃까지의 거리를 비교하여 거리가 더 작은 분류 항목을 선택하는 분류 전략이 $k = 2k' - 1$을 만족하는 기존의 k-최근접 이웃 분류 방법의 다수결 방법과 동일한 분류 결과를 낸다.

증명: 두 방법이 모든 경우의 수에 대해서 같은 분류 결과를 냄을 증명할 수 있다. 자세한 증명은 [6]에서 찾아볼 수 있다.

따름정리 1 (파라미터 추정방법과 다수결 방법의 동일성): 이진 분류 문제에서, 식 (1)의 모델과, 각 분류 항목의 k'번째 최근접 이웃까지의 거리를 이용하여 각 분류항목의 확률 밀도 을 최대 우도 방법으로 얻을 수 있다. 이 때, 이 추정된 확률 밀도 파라미터로 베이즈 분류한 결과가 을 만족하는 기존의 k-최근접 이웃 분류 방법의 다수결 방법과 동일한 분류 결과를 낸다.

증명: 각 분류 항목의 k'번째 최근접 이웃까지의 거리를 비교하였을때, 거리가 짧은 분류 항목이 더 큰 확률 밀도 추정값을 얻게 된다. 따라서, 추정값 $\hat{\lambda}_1, \hat{\lambda}_2$ 를 가지고 베이즈 분류를 하는 것은 k'번째 최근접 이웃까지의 거리가 짧은 분류 항목을 선택하는 분류와 항상 같다.

따름정리 2 (파라미터 추정방법과 다수결 방법의 동일성): 이진 분류 문제에서, 식 (1)의 모델과, 각 분류 항목의 첫 번째부터 k'번째 최근접 이웃들까지의 거리를 이용하여 각 분류항목의 확률 밀도 $\hat{\lambda}_1, \hat{\lambda}_2$ 을 최대 우도 방법으로 얻을 수 있다. 이 때, 이 추정된 확률 밀도 파라미터로 베이즈 분류한 결과가 $k = 2k' - 1$ 을 만족하는 기존의 k-최근접 이웃 분류 방법의 다수결 방법과 동일한 분류 결과를 낸다.

증명: 감마 분포에서 $p(u_{k'}|u_{k'-1}, \ldots, u_1, \lambda) = p(u_{k'}|\lambda)$ 를 만족하므로, 각 분류 항목의 k'번째보다 가까운 모든 최근접 이웃까지의 거리로 얻은 확률 밀도 추정값은 k'번째 최근접 이

웃의 거리만으로 추정한 확률 밀도와 항상 같게 된다. 이제 따름정리 1에 의해 이렇게 추정된 $\hat{\lambda}_1, \hat{\lambda}_2$ 를 이용한 베이즈 분류는 $k = 2k' - 1$ 을 만족하는 기존의 k-최근접 이웃 분류 방법의 다수결 방법과 동일한 분류 결과를 낸다.

정리 1과 따름정리 1, 2는 기존의 k-최근접 이웃 분류기의 다수결 방법이 추론을 통한 새로운 방법에서 어떤 위치에 있는지 이해할 수 있게 해 준다. 즉, 기존의 다수결 원칙을 추론에 의한 분류 문제로 볼 수 있는데, 최대 우도법을 이용한 확률 밀도의 추정과 이렇게 추정된 확률 밀도를 이용한 베이즈 분류로 k-최근접 이웃 분류기를 이해할 수 있다.

Ⅲ-2. 베이지안 추론

이전 장에서 본 것처럼 기존의 k-최근접 이웃 분류기를 식 (1)의 모델을 사용한 최대 우도 추정법으로 이해할 수 있다면, 자연스럽게 보다 좋은 추정을 위해 사전 추정을 이용한 베이지안 추론 방법의 적용을 생각해 볼 수 있다.

확률 밀도 λ 에 대한 사전 확률로 지수 분포 밀도 함수의 켤레 사전(conjugate prior) 확률 분포로서 λ 에 대한 지수 분포 밀도 함수를 고려할 수 있다.

$$p(\lambda) = b\exp(-b\lambda).$$

여기서 b는 사전 분포에 적용되는 하이퍼 파라미터이다. 이 사전 분포 밀도 함수와 최근접 이웃까지의 거리에 대한 모델인 식 (1)을 이용하여, 다음과 같은 닫힌 해로 제공되는 확률을 얻을 수 있다.

$$p(\lambda_1 > \lambda_2 | u_1, u_2) \quad (3)$$
$$= \sum_{m=0}^{k} \binom{2k+1}{m} \frac{(u_1+b)^m (u_2+b)^{2k+1-m}}{(u_1+u_2+2b)^{2k+1}}$$

단, $\binom{2k+1}{m}$ 는 ${}_{2k+1}C_m$ 를 나타낸다. 여기서, 각각의 분류 항목 데이터 중 k번째 최근접 데이터를 고려했으며, u_1, u_2 는 각 분류 항목의 k번째 최근접 데이터까지의 거리와 각 분류 항목의 데이터 개수로부터 구해졌다. 이렇게 계산된 확률값은 두 분류 항목의 k번째 최근접 데이터 거리가 주어졌을 때, 어떤 분류 항목의 모분포 밀도값이 클 것인가, 또한 어느 정도의 확실성을 가지고 클 것인가에 대한 정보를 제공한다. 사전 확률의 하이퍼파라미터 b 값을 조절하여 확실성 정도를 조절할 수 있다.

이렇게 만들어진 사후 확률값은 많은 정보를 제공한다. 우선 여기서 제공되는 확률값은 우리가 관심 있는 모분포의 직접 비교를 고려한다. 다시 말해, $p(\lambda_1 > \lambda_2 | u_1, u_2) = 0.7$ 일 때, 분류 항목 1로 분류하게 되면, 이 분류 결과가 베이즈 분류의 결과와 같을 확률이 0.7이라는 의미이다.

우리는 이 값을 적응 k-최근접 이웃(adaptive k-nearest neighbor) 분류에 이용할 수 있다. 즉, 새로운 최근접 이웃들을 얻을 때마다 계산 시간이 들어가는데, 반면 T. Cover 의 연구에 의하면 모분포가 비슷한 구간에서는 더 많은 최근접 이웃들을 얻을수록 성능은 좋아진다. 따라서 계산 시간과 분류 정확도가 서로 상쇄되고 어느 시점에서 타협을 이루어야 하는데 (tradeoff), 우리는 총 연산 시간이 동일할 때 더 높은 정확도를 얻거나, 혹은 같은 정확도를 낼 때 보다 적은 시간을 소비하는 알고리즘을 선호한다.

어느 정도 이상의 정확도를 얻을 확신이 있는 경우 더 이상의 최근접 이웃을 찾지 않고 분류를 행하게 된다. 이 때, 확신의 정도를 나타내는 기준(criterion)으로 식 (3)의 사후확률을 사용할 수 있다. 즉, 최근접 이웃들을 하나씩 구해 나가다 $p(\lambda_1 > \lambda_2 | u_1, u_2)$ 이나 $p(\lambda_1 < \lambda_2 | u_1, u_2)$ 이 미리 정해 놓은 기준값보다 큰 시점이 되면 더 이상의 최근접 이웃을 구하지 않고, 분류를 한다는 것이다. 앞의 확률이 기준값을 넘으면 분류 항목 1로, 뒤의 확률이 기준값을 넘으면 분류 항목 2로 분류하는 전략을 생각해 볼 수 있다.

Ⅳ. 최근접 이웃 정보를 이용한 분류 전략

앞 절에서는 모분포의 밀도 λ 를 최대 우도법을 통해 추정하여 분류에 이용하는 방법을 보았고 (k-최근접 이웃 분류법과 동일), 하나의 분류 항목의 밀도 λ 가 다른 분류 항목의 밀도보다 클 확률을 사전 확률과 베이지안 추론을 통해 얻을 수 있었다. 비슷한 모델과 접근 방법을 통해 분류에 사용될 수 있는 몇 가지 다양한 기준(criterion)을 만들 수 있다.

먼저, 우리의 모델은 최근접 이웃까지의 거리를 고려하는 식 (1)과 함께, 특정 반지름의 초구 내부에 포함되는 최근접 이웃의 수에 대한 분포인 포아쏭 분포를 같이 고려한다.

$$p(u_k|\lambda) = \frac{\lambda^k}{\Gamma(k)} \exp(-\lambda u_k)(u_k)^{k-1}, \quad (4)$$

$$p(k_u|\lambda) = \frac{(\lambda u)^{k_u}}{\Gamma(k_u+1)} \exp(-\lambda u). \quad (5)$$

식 (5)에서 k_u 는 $u = N\gamma d^D$ 일 때 거리 d 안에 있는 최근접 이웃의 수이다. 식 (4)는 식 (1)을 옮겨 놓은 식이고, 식 (5)는 k_u 개의 데이터가 반지름 d 안에 있을 확률을 나타낸다. 식 (5)는 체적 V 내부의 데이터 수 k 와 확률 밀도 λ 의 관계가 $\lambda = \mathbb{E}[k]/NV$ 와 같음을 고려하여 식의 타당성을 확인할 수 있다.

가장 먼저, 기존의 Wald의 Sequential Probability Ratio Test (SPRT) 기준을 이용할 수 있다. SPRT의 비율 테스트는 Wald test로도 알려져 있는데[1,4,12], 본 연구에서 다음과 같이 적용될 수 있다. 미리 정해진 두 개의 밀도 λ_+와 λ_-가 존재하며 $\lambda_+ > \lambda_-$를 만족 시키는데, 언제나 하나의 분류 항목은 λ_+를, 또 다른 분류 항목은 λ_-를 가지며, 두 분류 항목이 모두 λ_+를 가지거나 모두 λ_-를 가지지 않는다고 가정한다. 이러한 가정 하에, 어느 분류 항목이 더 큰 밀도 λ_+를 가지는지에 대한 테스트를 다음과 같이 α라는 확신값(confidence)을 두어 시행할 수 있다:

$$\frac{p(u_{1k}|\lambda_1=\lambda_+)p(u_{2k}|\lambda_2=\lambda_-)}{p(u_{1k}|\lambda_1=\lambda_-)p(u_{2k}|\lambda_2=\lambda_+)} > \alpha \quad \text{or} \quad < \frac{1}{\alpha}.$$

이 테스트는 식 (4)를 이용하여 다음과 같은 확신에 대한 기준(criterion)을 만들 수 있으며, 다음의 테스트와 동일하다는 것을 쉽게 유도할 수 있다.

$$|u_{1k} - u_{2k}| > z_V, \quad z_V = \frac{\log\alpha}{\lambda_+ - \lambda_-} \tag{6}$$

즉, 각 분류 항목의 k번째 최근접 이웃까지의 체적을 차, 혹은 거리의 차가 어느 이상이 되는지를 확인함으로써 확신성 테스트를 할 수 있게 된다. 비슷하게, SPRT를 최근접 이웃의 수에 적용을 시키면,

$$\frac{p(k_{1u}|\lambda_1=\lambda_+)p(k_{2u}|\lambda_2=\lambda_-)}{p(k_{1u}|\lambda_1=\lambda_-)p(k_{2u}|\lambda_2=\lambda_+)} > \alpha \quad \text{or} \quad < \frac{1}{\alpha},$$

아래와 같은 간단한 확신성 테스트의 기준을 얻게 된다.

$$|k_{1u} - k_{2u}| > z_N, \quad z_N = \frac{\log\alpha}{\log(\lambda_+/\lambda_-)} \tag{7}$$

또한, 같은 식의 변형을 통해서 또 하나의 기준을 얻을 수 있는데,

$$\max[(u_{1k} - u_{2(k+1)}), (u_{2k} - u_{1(k+1)})] > z_{V_2}, \tag{8}$$

이는 k개의 최근접 이웃을 포함하는 구는 u_k에 해당하는 체적을 갖기도 하지만, 한편, u_{k+1}의 체적보다 약간 작은 구도 k개의 최근접 이웃을 포함함을 고려하여 기준을 보다 충족시키기 어렵게 만들 수 있다.

우리는 식 (6)의 기준을 DV, 식 (7)의 기준을 DN, 그리고 식 (8)의 기준을 더 보수적인(conservative) DV라는 의미로 CDV라고 명명한다.

또한 베이지안 추론을 통해 이전 절에서 식 (4)의 모델을 통해 식 (3)과 같은 기준을 얻을 수 있었다.

$$p(\lambda_1 > \lambda_2|u_{1k}, u_{2k}) \tag{9}$$
$$= \sum_{m=0}^{k} \binom{2k+1}{m} \frac{(u_{1k}+b)^m (u_{2k}+b)^{2k+1-m}}{(u_{1k}+u_{2k}+2b)^{2k+1}}$$

비슷한 과정을 통해 식 (5)의 모델을 이용한 베이지안 추론을 통해 다음과 같은 기준을 얻을 수 있다.

$$p(\lambda_1 > \lambda_2|k_{1u}, k_{2u}) \tag{10}$$
$$= \frac{1}{2^{k_{1u}+k_{2u}+1}} \sum_{m=0}^{k_{1u}} \binom{k_{1u}+k_{2u}+1}{m}$$

우리는 베이지안 추론을 통해 얻어진 두 개의 기준을 각각 PV (식 (9)), PN (식 (10))이라고 명명한다.

이렇게 얻어진 다섯 개의 기준 DV, DV, CDV, PV, 그리고 PN은 모두 식 (4)와 식 (5)의 모델로부터 나온 기준들이다. 특별히 PV와 PN은 베이즈 분류기와 연관시켜 얼마나 정확히 베이즈 분류기와 같은 결과를 낼 것인가에 대한 확신을 준다는 데서 의미가 있다. 다음 절에서는 주어진 기준들을 사용한 적응 k-최근접 이웃 분류기가 어떻게 동작하는 지에 대한 실험들을 보여 준다.

V. 실험과 응용

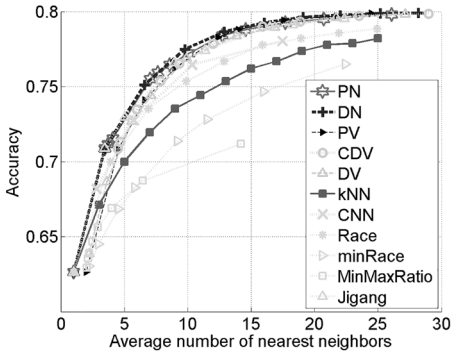


그림 2. 두 개의 서로 다른 확률 밀도값을 가지고 생성된 데이터를 이용한 실험. $\lambda_1 = 0.8, \lambda_2 = 0.2$. 베이즈 분류는 항상 분류 항목 1을 선택한다.

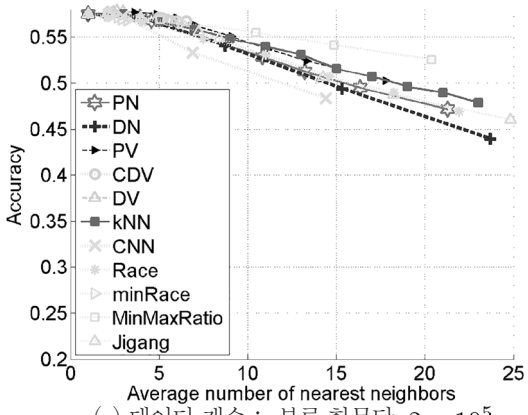
〈그림 2〉는 일정한 확률 밀도를 가지고 생성된 다른 두 분류 항목의 데이터에 대한 적응 k-최근접 이웃 분류기의 결과이다. 〈그림 2〉의 가로축은 확신성 테스트를 통과하는 기준값을 변화시켜 가면서 사용된 평균 최근접 이웃들의 수와 분류 성능을 나타낸 그래프이다. $\lambda_1 = 0.8, \lambda_2 = 0.2$을 사용 했으므로 최적 정확도인 베이즈 분류를 통한 정확도는 0.8이며, 확신도를 높여 사용된 평균 최근접 이웃의 수를 늘리면 모든 방법이 베이즈 분류의 정확도에 도달하는 것을 알 수 있다. 다른 기준들은 기존의 적응 k-최근접 이웃 분류기를 만들기 위한 다른 방법들이며, 그 분류 기준들은 대부분 직관적 동기에서 비롯되어왔다. CNN은 Consequent Nearest Neighbor의 약자로 한쪽 분

류 항목의 데이터가 연속으로 일정 기준 이상 나와야 항목을 결정하는 방법이고[8], Race는 어느 하나의 항목에 속하는 최근접 데이터 개수만 기준 이상이 되면 항목을 결정하는 방법이다. minRace는 모든 항목의 데이터 개수가 기준에 도달했을 때 다수결 원칙을 적용하는 방법이고[9][10][11], MinMaxRatio는 다른 항목의 최근접 이웃 수의 비율과 기준을 비교하는 방법이고, Jigang은 이 논문과 다른 확률 모델을 사용한 경우이다[13]. kNN은 모든 결과에서 k를 고정시키고, 다수결을 통해 분류를 한 방법이다.

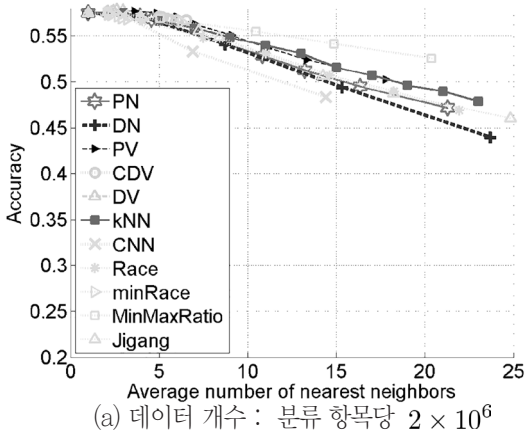
결과적으로 본 연구에서 제시된 식 (4)와 (5)를 모델로 이용하여 얻은 기준만 서로 비슷하면서 최고의 성능을 내고, 어떤 기존의 기준들은 일반적인 k-최근접 이웃 분류기의 결과보다도 안 좋은 성능을 보여준다.

〈그림 2〉의 결과는 최근접 이웃들이 나오는 지점들의 모분포가 일정할 경우 이러한 직관적인 전략들은 모두 최적의 결과를 내지 못 하지만, 최근접 이웃 정보의 이론적 지식에서 출발한 5개의 기준들을 이용한 전략들은 모두 비슷하게 가장 좋은 결과를 낸다는 것을 알 수 있다. 게다가, 서로 다른 다섯 개의 전략들이 비슷한 결과를 낸다는 것은, 이들이 내는 결과가 가장 최적의 분류 결과일 가능성이 크다는 것을 보여 준다.

〈그림 3〉은 두 개의 5차원 가우시안 모분포에서 데이터를 생성시켜 적응 k-최근접 이웃 분류한 결과이다. 단, (a)와 (b)는 같은 형태의 모분포를 사용했으나, 데이터의 개수를 달리 사용한 결과이다. 이 결과가 확실히 보여 주는 것은 이 연구의 이론적인 결과들이 데이터가 많아질 때 정확히 나오게 된다는 것이다. [6]은 실제 데이터에서 실행한 실험 결과를 포함하고 있다.



(a) 데이터 개수 : 분류 항목당 2×10^5



(a) 데이터 개수 : 분류 항목당 2×10^6

그림 3. 두 개의 가우시안 모분포에서 데이터를 생성시킨 후 적응 k-최근접 이웃 분류한 결과

Ⅵ. 다중 항목 분류로의 확장과 구현 방법

다중 분류 항목으로의 확장은 같은 맥락으로 계산 가능하지만(straightforward), 간단하지는 않다. SPRT의 확장은 각 분류 항목의 확실성 확률을 P_i라 할 때, DV는 다음과 같이 얻을 수 있고,

$$\log P_i = g^* k_i - \log\left(\sum_{c=1}^{C} \exp(g^* k_c)\right), \quad g^* = \log\frac{\lambda_+}{\lambda_-},$$

DN은 다음과 같이 얻을 수 있다.

$$\log P_i = -h^* u_i - \log\left(\sum_{c=1}^{C} \exp(-h^* u_c)\right), \quad h^* = \lambda_+ - \lambda_-.$$

여기서, 하나의 분류 항목만 더 큰 값인 λ_+를 가지고 있고, 나머지 분류 항목들은 모두 작은 값인 λ_-를 동일하게 가지고 있는 것을 가정한다. 이 때, 두 개 분류 항목의 경우 λ_+와 λ_- 값을 정확히 모르더라도 상관없이 알고리즘을 돌릴 수 있었던 것에 반해, 다중 항목 분류에서는 두 값의 비율이나 차이가 꼭 필요하다.

SPRT의 확장이 이렇게 λ_+와 λ_- 값을 아는 것이 알고리즘을 돌리는 데 필요하고, 큰 값인 λ_+를 제외한 나머지 모든 분류 항목이 동일한 작은 값 λ_-를 공유한다는 강한 가정을 사용한다면, 베이즈 추론을 이용할 때는, 여전히 각 분류 항목의 밀도값이 다를 수 있고, 특정 분류 항목의 밀도가 다른 모든 분류 항목들의 밀도보다 클 확률을 고려할 뿐이다. 이는 다중 항목 분류에서도 여전히 베이즈 분류와 같은 결과를 낼 확률과 동일하다[7].

이 확률은 데이터 D가 주어졌을 때,

$$\begin{aligned} P(\lambda_1 > \lambda_2, \ldots, \lambda_C | D) = 1 - P(\lambda_1 < \lambda_2 | D) - \\ \ldots - P(\lambda_1 < \lambda_C | D) + \\ \ldots + (-1)^{C-1} P(\lambda_1 < \lambda_2, \ldots, \lambda_C | D) \end{aligned}$$

와 같은 여러 확률 요소들의 조합으로 구해질 수 있는데, 각각

의 확률들은 PV의 경우

$$P(\lambda_1 < \lambda_{j_2}, \ldots, \lambda_{j_L} | u_1, \ldots, u_C) = \sum_{i_{j_2}=0}^{k_{j_2}} \cdots \sum_{i_{j_L}=0}^{k_{j_L}} \binom{k_1 + \sum_{c=2}^{L} i_{j_c}}{i_{j_2}, \cdots, i_{j_L}} \frac{u_1^{k_1+1} \prod_{c=2}^{L} u_{j_c}^{i_{j_c}}}{(u_1 + \sum_{c=2}^{L} u_{j_c})^{k_1 + \sum_{c=2}^{L} i_{j_c} + 1}},$$

PN의 경우

$$P(\lambda_1 < \lambda_{j_2}, \ldots, \lambda_{j_L} | k_1, \ldots, k_C) = \sum_{i_{j_2}=0}^{k_{j_2}} \cdots \sum_{i_{j_L}=0}^{k_{j_L}} \frac{1}{L^{(k_1+1+\sum_{c=2}^{L}(k_{j_c}-i_{j_c}))}} \binom{k_1 + \sum_{c=2}^{L}(k_{j_c} - i_{j_c})}{k_{j_2} - i_{j_2}, \cdots, k_{j_L} - i_{j_L}}$$

와 같이 구할 수 있다.

추론 결과로 나온 확률을 구하기 위한 구현은 많은 수의 분류 항목 개수와 많은 수의 최근접 이웃 개수의 경우에 많은 시간을 필요로 한다. 일반적인 분류 항목 개수에 대해 구현하기 위해서 PN과 PV의 다중 합(summation)은 재귀(recursion)를 통해 구현될 수 있다. 다항 계수(multinomial coefficient)를 계산하는 데 있어서는 모든 상수를 따로 계산하지 않고 증가를 통해 계산함으로써 시간을 많이 절약할 수 있다. 다항 계수의 다음 성질을 이용할 수 있다.

$$\binom{N}{k_1, \ldots, k_i, \ldots, k_C} = \binom{N-1}{k_1, \ldots, k_i - 1, \ldots, k_C} \frac{N}{k_i}$$

Ⅶ. 결 론

본 리뷰 논문에서는 최근접 이웃 정보의 이론적 특성을 이용한 최신 분류 방법에 관해 정리하여 소개했다. 기존에 일상적으로 사용해 온 다수결을 통한 k-최근접 이웃 분류법은 이런 측면에서 봤을 때 상당히 우수한 방법이다. 하지만, SPRT와 베이지안 추론을 통해 기존의 k-최근접 이웃 분류법보다 성능/시간 상호 득실(tradeoff) 상황에서 가장 효과적인 성능을 내는 전략들을 만들어 낼 수 있었다. 적용된 이론이 데이터가 많아 최근접 데이터들 사이에 같은 밀도 함수값을 공유하는 상황을 가정했으므로, 소개된 방법은 특별히 데이터가 많은 상황에서 아주 명확하게 두각을 나타날 것으로 기대된다.

Acknowlegement

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신 방송연구개발사업[14-824-09-014, 인간 수준의 평생 기계학습SW 기초 연구(기계학습연구센터), 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단(No. NRF-2011-0009963, No. NRF-2012-R1A2A4A01008475), 2014년 두뇌한국21플러스 사업 및 연구재단 일반연구자 지원사업 NRF-2012R1A1A2007881의 지원을 받아 수행하였음.

참 고 문 헌

[1] Bogacz,R.,Brown,E.,Moehlis,J.,Holmes,P.,& Cohen,J. D. (2006) The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4), 700-765

[2] Cover, T. (1967) Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory, 14*(1), 50-55.

[3] Cover, T., & Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27.

[4] Dragalin, V. P., Tertakovsky, A. G., & Veeravalli, V. V. (1999) Multihypothesis sequential probability ratio tests. part i: asymptotic optimality. *IEEE Transactions on Information Technology, 45*, 2448-61.

[5] Leonenko,N.,Pronzato,L.,& Savani,V. (2008) A class of Renyi information estimators for multidimensional densities. *Annals of Statistics, 36*, 2153-2182.

[6] Noh, Y.K., Park, F.C., & Lee, D.D. (2012) Diffusion Decision Making for Adaptive k-Nearest Neighbor Classification, *Advances in Neural Information Processing Systems 25*

[7] Noh, Y.K., Park, F.C., & Lee, D.D. (2013) k-Nearest Neighbor Classification Algorithm for Multiple Choice Sequential Sampling, *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society*

[8] Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A. N., Manolopoulos, Y., & Welzer-Druzovec, T. (2007) Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors. *In Proceedings of the 11th east European conference on advances in databases and information systems* (pp. 66-82).

[9] Smith, P. L., & Vickers, D. (1988) The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology, 32*, 135-168.

[10] Usher, M., & McClelland, J.L. (2001) The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review, 108*(3), 550-592.

[11] Vickers, D. (1970) Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*, 37-58.

[12] Wald, A., & Wolfowitz, J. (1948) Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics, 19*, 326-339.

[13] Wang, J., Neskovic, P., & Cooper, L. N. (2006) Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Pattern Recognition, 39*(3):417-423.

약 력

노 영 균

1998년 포항공과대학교 물리학과 졸업(이학사)
2011년 서울대학교 협동과정 인지과학 졸업(공학박사)
2007년~2012년 펜실베니아 대학 방문연구원
2011년~2013년 서울대학교 기계항공공학부 포닥
2013년~현재 한국과학기술원 전산학과 연구교수
관심분야: 기계학습, 거리학습, 차원축소, 대용량 데이터 분석

김 기 응

1995년 KAIST 전산학과 학사
1998년 브라운대학교 전산학과 석사
2001년 브라운대학교 전산학과 박사
2001년~2003년 삼성 SDS 책임연구원
2004년~2006년 삼성종합기술원 책임연구원
2006년~2012년 KAIST 전산학과 조교수
2012년~현재 KAIST 전산학과 부교수
관심분야: 인공지능, 기계학습, 강화학습

이 태 훈

2009년 고려대학교 공학사
2012년 고려대학교 공학석사
2012년~현재 서울대학교 전기정보공학부 박사과정
관심분야: 대용량 바이오 컴퓨팅, 확률 통계 이론

윤 성 로

1996년 서울대학교 공학사
2002년 Stanford University 공학석사
2006년 Stanford University 공학박사
2006년 Stanford University 박사후 연구원
2006년~2007년 미국 인텔 선임연구원
2007년~2012년 고려대학교 전기전자전파공학부 조교수
2012년~현재 서울대학교 전기정보공학부 부교수
관심분야: 기계학습, 생체정보, 빅데이터 분석, 고성능 컴퓨팅 및 스토리지

Daniel D. Lee

1990년 하버드 물리학과 졸업(이학사)
1995년 MIT 물리학과 졸업(이학박사)
1995년~2001년 Research and Development Arm of Lucent Technologies
2001년~현재 펜실베니아 대학 전자시스템학과 교수
관심분야: 로보틱스, 기계학습