The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 401–410

DOI: https://doi.org/10.5351/KJAS.2021.34.3.401
Print ISSN1225-066X / Online ISSN 2383-5818

Random projection ensemble adaptive nearest neighbor classification

Jongkyeong Kang[a], Myoungshic Jhun[1,b]

[a]Department of Statistics, Korea University;
[b]Department of Applied Mathematics and Statistics, The State University of New York Korea

Abstract

Popular in discriminant classification analysis, k-nearest neighbor classification methods have limitations that do not reflect the local characteristic of the data, considering only the number of fixed neighbors. Considering the local structure of the data, the adaptive nearest neighbor method has been developed to select the number of neighbors. In the analysis of high-dimensional data, it is common to perform dimension reduction such as random projection techniques before using k-nearest neighbor classification. Recently, an ensemble technique has been developed that carefully combines the results of such random classifiers and makes final assignments by voting. In this paper, we propose a novel discriminant classification technique that combines adaptive nearest neighbor methods with random projection ensemble techniques for analysis on high-dimensional data. Through simulation and real-world data analyses, we confirm that the proposed method outperforms in terms of classification accuracy compared to the previously developed methods.

Keywords: adaptive nearest neighbor, classification, high-dimensional data, K-nearest neighbor, random projection

1. 서론

판별분류분석은 라벨링된 훈련자료를 기반으로 새로운 개체를 둘 이상의 집단 중 하나에 할당하는 통계적 기법으로, 행동 인식 (Mohino-Herranz 등, 2019), 의료 진단 (Chang과 Kwon, 2008), 신용 평가 (Kim과 Kim, 2013), 자연어 처리 (Kang 등, 2015) 등 수많은 응용 분야에서 널리 활용되고 있다. k-최근접 이웃 분류(k-nearest neighbor classification; KNN)는 Fix와 Hodges (1951)에 의해 처음 소개된 비모수적 분류기법으로, k개의 가장 가까운 훈련 개체의 라벨에 대한 다수결에 따라 가장 빈도가 높은 집단으로 시험 개체를 분류한다. KNN은 단순하면서도 직관적일 뿐만 아니라, 이웃의 수 k가 무한히 커지면서 동시에 표본의 크기 n 에 대한 비 k/n이 0으로 갈 때, KNN 분류의 위험이 베이즈 위험으로 수렴하는 매력적인 특성으로 널리 이용되고 있다 (Devroye 등, 1996). 그러나 KNN은 전체 자료에 대해 고정된 이웃의 수 k를 선택하기 때문에, 각 개체가 갖는 국소적 특성을 고려하지 않는다. 이러한 KNN 방법은 k개의 가장 가까운 이웃 관찰값에 근거하여 통계적 분석을 수행하는데, 이 때 적절한 k의 선택은 분석의 성능을 결정하는 매우 중요한 요소가 된다. 너무 작은 k는 과대적합을, 반면에 너무 큰 k의 사용은 과소적합을 초래한다. 따라서, 고정된 k의 사용은 분석 대상 자료의

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2020R1F1A1A01061746) for M. Jhun.
[1] Corresponding Author: Department of Applied Mathematics and Statistics, The State University of New York Korea, 119 Songdo Moonhwa-Ro Incheon, 21985, Korea. E-mail: myoungshic.jhun@stonybrook.edu

Published 30 June 2021 / journal homepage: http://kjas.or.kr

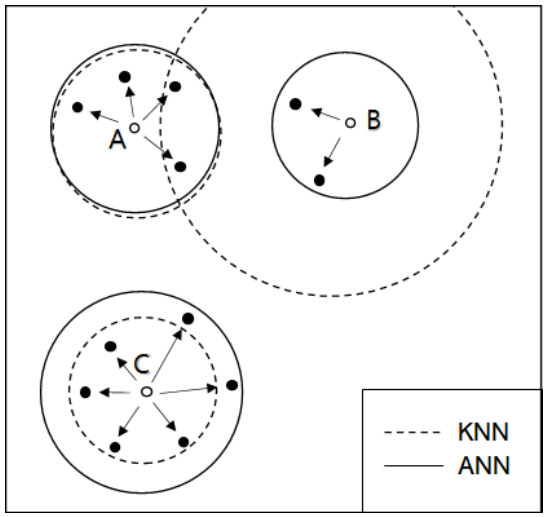

Figure 1: *Illustration of KNN and ANN classification.*

국소적 상황이나 집단 분포의 치우침 등을 반영할 수 없다는 단점을 지니게 된다 (Hastie와 Tibshirani, 1996). 이에 각 개체에 따라 소속 집단의 결정에 사용되는 이웃의 수를 변화시키는 방법에 대해 Friedman (1994), Hastie와 Tibshirani (1996) 등의 연구가 이루어져 왔다. Jhun과 Choi (2009)는 분류 대상 이웃 선정에 있어 자료의 구조나 밀도를 반영하여 선택하는 이웃 수를 적응적으로 바꾸는 적응 최근접 이웃(adaptive nearest neighbor; ANN) 판별분류 기법을 제안하였다.

한편 현대 많은 분류 문제는 설명 변수의 수 p가 훈련자료의 크기 n과 비슷하거나 더 큰 고차원 자료를 다루고 있다. 다른 비모수 방법들과 마찬가지로 KNN 역시 일반적으로 고차원에서 저조한 성능을 보인다는 '차원의 저주' 효과가 발생한다 (Hall 등, 2008). 랜덤 스케치 또는 랜덤 투영(random projection)은 고차원 자료를 저차원에서 다루기 위해 이용되는 대표적인 차원 축약 방법으로, 자료의 선형변환을 통해 차원 축약을 가져온다 (Bingham과 Mannila, 2001; Shasha와 Zhu, 2004). Kang과 Jhun (2020a)은 대용량 자료에서의 커널 릿지 회귀에서 분할정복 기법과 랜덤스케치 기법을 결합하였으며, Kang과 Jhun (2020b)은 랜덤스케치 기법을 활용하여 재생성 힐버트 공간에서의 변수선택에 적용하였다. Cannnings와 Samworth (2017)는 이원 분류 문제에 있어 p차원의 설명변수들을 $d < p$ 차원 공간에 무작위로 여러 번 투영한 후 각각 분류한 뒤, 앙상블 기법을 적용하여 이들의 분류 결과를 자료에 기반한 투표를 통해 집계한 후 최종 할당을 결정하는 방법을 제안했다.

본 논문에서는 고차원 자료에서의 판별분류분석을 위해 랜덤 투영 앙상블 적응 최근접 이웃 분류 기법 (random projection ensemble adaptive nearest neighbor; RPEANN)을 제안하고자 한다. 본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존의 ANN에 관해 간략하게 소개하였다. 3장에서는 제안 방법인 랜덤 투영 앙상블을 활용한 ANN 기법에 대해 소개하였으며, 4장에서는 제안된 방법의 유용성을 모의실험과 실제 사례 분석을 통해 확인하였다.

2. 적응 최근접 이웃 분류

소속 집단 $c_1, \ldots, c_g$을 나타내는 반응변수 $y \in G := \{c_1, \ldots, c_g\}$와 p차원 연속형 설명변수 $\mathbf{x} \in \mathbb{R}^p$로 이루어진 n개의 개체에 대한 자료 $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$를 고려하자. 새로운 개체 $\mathbf{x}$와 가장 가까운 이웃들을 정하기 위해서는 '가까움'을 측정하는 측도가 필요하며, $\mathbf{x}$와 $\mathbf{x}_i$의 거리를 $d(\mathbf{x}, \mathbf{x}_i)$라고 하자. 이러한 측도로 유클리디안

Algorithm 1 적응 최근접 이웃 분류

[단계 1] 주어진 점 $\mathbf{x}$에 대하여 거리 기준 $d(\ldots, \mathbf{x})$에 따라 $\|\mathbf{x}_{(1)}, \mathbf{x}\| \le \cdots \le \|\mathbf{x}_{(n)}, \mathbf{x}\|$가 되도록 오름차순으로 크기순으로 정렬하여 $(\mathbf{x}_{(1)}, y_{(1)}), \ldots, (\mathbf{x}_{(n)}, y_{(n)})$을 만든다.

[단계 2] 주어진 점 $\mathbf{x}$로부터 $\mathbf{x}_{(i)}$까지의 거리 $d(\mathbf{x}_{(i)}, \mathbf{x})$에 조정계수 δ를 더해 조정된 거리 $d^*(\mathbf{x}_{(i)}, \mathbf{x}) = d(\mathbf{x}_{(i)}, \mathbf{x}) + \delta$를 정의한다.

[단계 3] 제어계수 q를 설정한 다음, $\mathbf{x}$의 이웃으로

$$N_{\mathbf{x}} = \{(\mathbf{x}_{(i)}, y_{(i)}) : d^*(\mathbf{x}_{(i)}, \mathbf{x}) \le q d^*(\mathbf{x}_{(1)}, \mathbf{x})\}$$

를 정한다.

[단계 4] $\mathbf{x}$의 소속 집단으로 $c^* = C_n(\mathbf{x}) = \text{argmax}_{c_j \in G} \sum_{(\mathbf{x}_{(i)}, y_{(i)}) \in N_{\mathbf{x}}} I(c = c_j)$을 할당한다.

거리, 마할라노비스 거리, 상관계수 등 다양한 방법들을 사용할 수 있다. ANN 방법은 새로운 개체 $\mathbf{x}$와 가장 가까운 개체와의 거리에 대한 다른 개체들과의 거리의 비를 이용하여 이웃집단을 구성한다. 이러한 비는 $\mathbf{x}$가 속한 국소적 위치의 분포에 따라 달라지므로, 분류에 사용되는 이웃의 개수도 달라지게 된다.

Figure 1은 ANN의 가능성을 제시하고 있다. 개체 A, B, C에 대해서 고정된 $k = 4$개의 이웃을 사용하는 KNN은 점선내의 개체들을 이웃으로 택한다. 반면에, 실선으로 표현되는 ANN은 개체 A는 4개의 이웃을 개체 B는 2개의 이웃을 그리고 개체 C는 6개의 이웃을 택한다. 즉 ANN은 자료의 국소적 특징을 반영하여 이웃의 개수를 결정함으로써 자료의 국소적 상황을 보다 타당하게 고려한다. Algorithm 1은 ANN 판별분류기 $C_n : \mathbb{R}^p \times G \to G$을 설계하는 알고리즘을 보여준다.

Algorithm 1은 주어진 점 $\mathbf{x}$로부터 가장 가까운 개체부터 차례로 이웃으로 선택하는 과정에서 다음 이웃까지의 거리가 일정 수준 이상이 될 경우 이웃으로의 선택을 중단하는 방법으로 자료의 국소적 상황을 반영한다. [단계 3]에서 제어 계수 q는 $\mathbf{x}$의 이웃으로서의 선택 한계를 결정하는 절단값(cutoff point)으로 q가 클 수록 더 많은 이웃을 갖게 된다. 이러한 방법으로 이웃을 결정함에 있어 만약 점 $\mathbf{x}$로부터 가장 가까운 거리 $d(\mathbf{x}_{(1)}, \mathbf{x})$가 0에 매우 가깝다면 임의의 다른 점과의 거리의 비 $d(\mathbf{x}_{(i)}, \mathbf{x})/d(\mathbf{x}_{(1)}, \mathbf{x})$가 무한히 커지게 되며, 이 경우 가장 가까운 점 이외에 다른 이웃을 선택하지 못할 수 있다. 이에 대한 보완책으로 [단계 2]와 같이 거리의 순위에는 아무런 영향 없는 조정계수 δ를 더한 조정된 거리 $d^*(\mathbf{x}_{(i)}, \mathbf{x}) = d(\mathbf{x}_{(i)}, \mathbf{x}) + \delta$를 사용하는 것이 바람직하며, 실제 자료분석에서는 $d(\mathbf{x}_{(i)}, \mathbf{x})$들의 중위수를 조정계수 δ로 사용할 수 있다. 또한 δ가 임의로 설정되어도, 제어계수 q의 면밀한 선택을 통해 포함하는 이웃의 수의 조율이 가능하다.

3. 랜덤 투영 앙상블 적응 최근접 분류

고차원 자료에서의 랜덤 투영의 이용은 Johnson-Lindenstrauss 보조정리 (Johnson과 Lindenstrauss, 1984)에 제시되어 있는데, 이는 벡터 공간의 점들이 충분히 높은 차원이라면 점들 사이의 거리를 대략적으로 보존하는 방식으로 적절한 저차원 공간에 투사될 수 있다고 말한다. 즉, 주어진 $0 < \epsilon < 1$과 $\mathbb{R}^p$에 있는 n개의 p차원 벡터 $\mathbf{x}_1, \ldots, \mathbf{x}_n$, 그리고 $d > 8 \log n/\epsilon^2$에 대하여, 모든 $i, j = 1, \ldots, n$에 대해

$$(1 - \epsilon)\|bx_i - \mathbf{x}_j\|^2 \le \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 \le (1 + \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2,$$

을 만족하는 선형 변환 $f : \mathbb{R}^p \to \mathbb{R}^d$이 존재한다. 여기에서 흥미로운 사실은 d의 하한이 $\mathbf{x}$의 차원 p가 아닌 표본의 크기 n 에 의존한다는 것이다. 차원 p가 $\log n$에 비해 큰 경우, 자료를 저차원 공간에 무작위로 투영하고 투영된 자료에 대한 통계 절차를 실행하여 잠재적으로 상당한 계산 절감 효과를 얻을 수 있으며, 이와 유사하거나 심지어 개선된 통계 성능을 달성할 수 있다.

투영 차원 $d \le p$에 대해 $AA^\intercal = I_d$을 만족하는 $A \in \mathbb{R}^{d \times p}$를 투영행렬(projection matrix)이라고 한다. 여기서

Algorithm 2 랜덤 투영 앙상블 적응 최근접 이웃 분류

[단계 1] $\mathscr{A}$로부터 독립적으로 추출한 m개의 투영행렬 $A_1, \ldots, A_m$을 추출한다.

[단계 2] Algorithm 1에서 설계한 ANN 판별분류기 C_n를 이용하여 m개의 판별분류기를 $C_n^{A_1}, \ldots, C_n^{A_m}$를 다음과 같이 설계한다

$$C_n^{A_i}(\mathbf{z}^i) = C_n(A_i\mathbf{x}).$$

여기서 $\mathbf{z}^i = A_i\mathbf{x}$ 이다.

[단계 3] $\mathbf{x}$의 소속 집단으로

$$c^* = C_n^{RP}(\mathbf{x}) = \text{argmax}_{c_j \in G} \sum_{k=1}^{m} I(C_n(A_k\mathbf{x}) = c_j),$$

을 할당한다.

I_d는 d차원 단위행렬이다. $\mathscr{A} = \{A \in \mathbb{R}^{d\times p} : AA^\intercal = I_d\}$를 모든 투영행렬들의 집합이라고 하자. 투영행렬 $A \in \mathscr{A}$가 주어져 있을 때, $i = 1, \ldots, n$에 대해 $\mathbf{z}_i = A\mathbf{x}_i$로 정의하자. 2장에서 살펴본 ANN 분류기 C_n을 이용하여 투영행렬 A에 의해 랜덤 투영된 자료 $\{(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_n, y_n)\}$로 부터 설계한 ANN 판별분류기 $C_n^A(\mathbf{z})$는 다음과 같이 정의할 수 있다.

$$C_n^A(\mathbf{z}) = C_n(A\mathbf{x}).$$

여기서 분류기를 설계하는 데 이용되지 않은 새로운 점 $\mathbf{x}$ 역시 동일한 투영행렬 A에 의해 $\mathbf{z} = A\mathbf{x}$의 형태로 투영되어야 한다. 이제 앙상블 기법을 적용하기 위해 $\mathscr{A}$로부터 독립적으로 m개의 투영행렬 $A_1, \ldots, A_m$을 추출했다고 하자. 여기서 투영행렬들은 자료와는 독립이다. 그러면 우리는 m개의 투영된 자료에 대한 ANN 분류기 $C_n^{A_1}, \ldots, C_n^{A_m}$를 설계할 수 있다. 본 연구에서 제안하는 랜덤 투영 앙상블 적응 최근접 이웃(random projection ensemble adaptive nearest neighbor; RPEANN) 판별분류기 $C_n^{RP} : \mathbb{R}^p \times G \to G$을 설계하는 알고리즘은 Algorithm 2에 나타나 있다.

랜덤 투영 행렬 A를 하르 측도(Haar measure)로 부터 생성하기 위해 우선 각 성분이 표준정규분포를 따르는 행렬 $B \in \mathbb{R}^{d\times p}$를 우선 생성한 후, $B^\intercal$의 특잇값분해(singular value decomposition)의 좌특이벡터(left singular vector)들의 행렬의 전치를 취하는 방법을 생각할 수 있다. 이 경우 계산 비용은 $O(pd^2)$가 된다 (Trefethen과 Bau, 1997). 그러나 실제 적용에 있어서 행렬 A의 직교성은 모형의 성능에 크게 영향을 주지 않는다 (Cannnings와 Samworth, 2017). 본 논문의 모의실험 및 실제자료의 분석에서는 랜덤 투영 행렬 A의 각 성분이 평균이 0이고 분산이 $1/p$인 정규분포를 따르는 성분을 갖도록 생성하였으며, 이 경우 계산 비용은 $O(pd)$로 줄게 된다. 또한 d가 너무 크게 설정되는 경우 고차원 자료의 분석에서 KNN이나 ANN이 갖는 차원의 저주를 답습하게 된다. 따라서 투영 차원 d는 실제 두 점 사이의 거리를 너무 왜곡하지 않으면서도 충분히 작은 값으로 설정하는 것이 바람직하다.

Algorithm 2에서 확인할 수 있듯이, 랜덤 투영 앙상블 기반 방법은 m개의 분류기에서의 분류 결과를 종합하여 다수결을 통해 최종 분류를 할당하게 된다. 즉, 분류기의 수 m이 크면 클수록 다수결에 의한 결정이 일관되어 안정적인 성능을 보여주지만, 본질적으로 기존의 방법을 m번 반복하여 수행하게 되므로 그만큼 계산의 소요가 증가하게 된다. 따라서 d의 설정에 있어서와 마찬가지로, m 역시 안정적인 성능을 보여주면서 충분히 작은 값으로 설정하는 것이 바람직하다. 특정 점에서 정분류하는 비율이 모두 π로 동일한 m개의 분류기가 있다고 가정해보자. 그러면, $\alpha > 0$와 $B \sim \text{Binom}(m, \pi)$에 대해

$$m = \text{argmin}_m \left[P(B > \frac{m}{2}) > 1 - \alpha\right],$$

Table 1: Possible choice of m with respect to the probability of correct classification for π

m		π					
		0.8	0.75	0.7	0.65	0.6	0.55
α	0.10	5	9	13	23	49	183
	0.05	8	13	19	35	75	289

를 만족하는 m을 고려할 수 있다. 즉, 그 점에 대해 m개의 분류기에 의한 다수결의 결과가 정분류할 확률이 $1-\alpha$보다 크게 될 것으로 기대되는 m을 선택하는 것이다.

Table 1은 π값에 따라 정분류할 확률이 $1-\alpha$가 되는 최소의 m의 값을 보여주고 있다. 예를 들어, $\pi = 0.8$의 경우 $m \geq 8$개의 분류기에 대한 다수결의 결과가 정분류할 확률은 0.95 이상이며, $\pi = 0.7$의 경우는, $m \geq 19$개의 분류기에 의한 다수결의 결과가 정분류할 확률은 0.95 이상이다. 실제로는 π가 알려져 있지 않으며, 각 시험 지점에서의 π의 값은 모두 다르다. 또한 $\pi < 0.5$인 지점의 경우 m이 커짐에 따라 오히려 정분류할 확률은 작아진다. 4장에서 기술될 모의실험 및 실제자료 분석에서는 다양한 m에 대해 실험을 하였으며, 본 제안 방법의 성능이 m의 선택에 있어서 강건함을 확인하였다.

한편 Algorithm 2의 [단계 1]과 [단계 2]는 각 $k = 1, \ldots, m$에 있어 독립적으로 이루어지며, 다중 프로세서 기반의 환경에서 병렬 분산 계산을 통해 효율성을 증대 시킬 수 있다.

4. 모의실험 및 실제 자료 분석

이 장에서는 제안된 랜덤 투영 앙상블 적응 최근접 이웃 분류 기법의 효율성을 기존의 방법인 KNN과 ANN, 그리고 랜덤 투영 앙상블 KNN (random projection ensemble KNN; RPEKNN)과 비교하기 위해 모의실험을 수행하고 나아가 실제 자료의 분석을 통해 활용 가능성을 살펴보았다. KNN 및 RPEKNN에서의 이웃의 개수에 대한 모수인 k는 1부터 30까지를 고려하였고, ANN 및 제안방법인 RPEKNN에서의 제어계수 q는 1부터 2까지 0.03씩 등간격으로 고려하였으며, 이들 모수의 선택은 교차타당성(leave-one-out cross validation) 방법을 통해 이루어졌다.

4.1. 모의실험

예제 1: 이원분류 문제를 고려하였다. $p \times p$ 행렬 Σ의 대각원소는 1이고, 비대각원소는 0.5 라고 하자. 집단 1(c_1)은 평균이 $\mu_1 = (1, \ldots, 1)^\intercal$이고 공분산행렬이 $\Sigma_1 = 1/2\Sigma$ 인 정규분포 $N(\mu_1, \Sigma_1)$부터 추출하였고, 집단 2(c_2)는 평균이 $\mu_2 = -\mu_1$ 이고 공분산행렬이 $\Sigma_2 = 2\Sigma$인 정규분포 $N(\mu_2, \Sigma_2)$로부터 추출하였다. 집단 2의 분산이 집단 1의 분산에 비해 4배가 크므로, 집단 2에 속한 일부 표본들이 집단 1의 표본들과 가깝게 생성될 가능성이 있게 설계하였다. 두 집단의 표본 크기는 각각 $n_1 = n_2 = 100$으로 정하였다. 랜덤 투영 앙상블 기반 방법인 RPEKNN 과 제안 방법인 RPEANN 분류기를 설계하기 위해서 투영 차원 d와 앙상블에 이용되는 분류기의 수 m을 결정해야 한다. 본 모의실험에서는 $d = (2, 5, 10, 20)$을, $m = (1, 10, 30, 60)$을 고려하였다. 여기서 $m = 1$일 때의 방법은 랜덤 투영한 자료에 대한 분류기(random projection classifier)를 의미한다. 시험 자료는 훈련자료와 동일한 집단으로부터 크기가 100이 되도록 생성하였다.

Table 2는 예제 1의 $p = 100$에서의 실험을 독립적으로 100회 반복한 결과 평균 정분류율을 나타내고 있다. Table 2에서 아래첨자는 100번 반복 시행에서의 정분류율의 표준편차를 나타낸다. ANN이 KNN에 비해 정분류율 측면에서 더 높은 성능을 보여줬다. KNN은 $d = 5$ 이상의 랜덤 투영 행렬에 의해 투영된 자료에 대한 정분류율이 기본적인 KNN의 정분류율보다 높았으며, RPEKNN은 모든 경우에서 기본적인 KNN보다 정분류율이 높았다. 제안 방법인 RPEANN은 $d = 2$ 또는 $d = 5$, 그리고 $m \geq 30$인 상황에서 기존의 ANN

Table 2: Classification accuracies for example 1

RPEKNN (KNN: $0.839_{0.037}$)					RPEANN (ANN: $0.888_{0.035}$)				
m	d				m	d			
	2	5	10	20		2	5	10	20
1	$0.833_{0.064}$	$0.869_{0.043}$	$0.865_{0.037}$	$0.852_{0.040}$	1	$0.836_{0.072}$	$0.875_{0.044}$	$0.875_{0.038}$	$0.866_{0.038}$
10	$0.880_{0.035}$	$0.879_{0.034}$	$0.868_{0.036}$	$0.856_{0.035}$	10	$0.894_{0.037}$	$0.886_{0.034}$	$0.875_{0.037}$	$0.872_{0.035}$
30	$0.898_{0.033}$	$0.886_{0.034}$	$0.872_{0.034}$	$0.861_{0.035}$	30	$0.910_{0.034}$	$0.905_{0.035}$	$0.881_{0.035}$	$0.875_{0.035}$
60	$0.901_{0.031}$	$0.890_{0.033}$	$0.872_{0.033}$	$0.863_{0.035}$	60	$0.913_{0.034}$	$0.904_{0.034}$	$0.893_{0.034}$	$0.887_{0.034}$

Table 3: Confusion matrix of KNN, ANN, RPEKNN, and RPEANN for example 1

		KNN		ANN		RPEKNN		RPEANN	
		c_1	c_2	c_1	c_2	c_1	c_2	c_1	c_2
Actual	c_1	49.88	0.10	49.8	0.18	49.45	0.53	49.27	0.71
	c_2	15.66	34.36	10.77	39.25	9.28	40.74	9.01	43.01

보다 평균적으로 더 높은 정분류율을 보여줬다. RPEKNN과 RPEANN 모두 $d = 2, k \geq 30$ 이상에서 가장 정분류율이 높았으며, RPEANN이 RPEKNN에 비해 정분류율의 측면에서 보다 우수한 성능을 보였다.

Table 3은 예제 1의 실험 결과에 대한 KNN과 ANN 그리고 $m = 30, d = 2$에서의 RPEKNN과 RPEANN의 혼동행렬(confusion matrix)을 보여준다. 혼동행렬의 각 성분은 100번 독립 반복 시행에서의 각 집단별 분류 결과의 평균값이다. KNN의 경우 집단 1을 집단 2로 오분류하는 경우는 거의 없었지만, 집단 2를 집단 1로 오분류한 경우가 다수 있었다. 랜덤 투영 앙상블 기법을 사용한 RPEKNN과 RPEANN은 집단 1을 집단 2로 오분류하는 경우가 조금 늘었지만, KNN과 ANN에 비해 집단 2를 집단 1로 오분류하는 경우를 상대적으로 많이 줄임을 통해, 전반적인 정분류율이 높아짐을 확인할 수 있다.

예제 2: 예제 1에 새로운 집단 3(c_3)을 추가한 삼원분류 문제를 고려하였다. 집단 1과 집단 2는 예제 1과 동일하게 추출하였으며, 새로운 집단 3은 평균이 $\mu_3 = (0, \ldots, 0)^\intercal$이고 공분산행렬이 $\Sigma_3 = \Sigma$ 인 정규분포 $N(\mu_3, \Sigma_3)$부터 추출하였다. 집단 1과 집단 2 사이에 위치한 집단 3을 추가하여 분류에 혼동을 가하도록 설계하였다. 세 집단의 표본 크기는 각각 $n_1 = n_2 = n_3 = 100$으로 정하였다. 이외의 나머지 실험 설계는 예제 1과 같다.

Table 4는 예제 2의 $p = 100$에서의 실험 결과를 독립적으로 100회 반복한 결과 평균 정분류율을 나타내고 있으며, 전반적인 실험 결과는 예제 1과 유사하다. Table 5는 예제 2의 실험 결과에 대한 KNN과 ANN 그리고 $m = 30, d = 2$에서의 RPEKNN과 RPEANN의 혼동행렬을 나타내고 있으며, 각 성분은 100번 독립 반복 시행에서의 각 집단별 분류 결과의 평균값이다. 예제 2에서는 집단 1을 집단 2로 오분류하는 경우는 모든 분류방법에서 없었으며, 집단 1은 전반적으로 정분류율이 높았다. 그러나 집단 2와 집단 3에 대한 분류에 있어서는 기존의 KNN과 ANN에 비해, 랜덤 투영 앙상블 기반 방법들의 정분류율이 유의하게 높았다. 랜덤 투영 앙상블 기반 방법간의 비교에 있어서도 제안 방법인 RPEANN이 RPEKNN에 비해 평균적으로 집단 3을 더 정확하게 분류함을 확인하였다.

4.2. 실제 자료 분석

제안 방법의 실제 자료 분석에의 활용 가능성을 살펴보기 위해 웨어러블 생리학적 측정장치를 이용한 행동 인식 자료 (Mohino-Herranz 등, 2019)를 분석하였다. Mohino-Herranz 등 (2019)은 웨어러블 계측기와 결합된 센서화된 의복을 사용하여 심전도(electrocardiogram; ECG), 흉부 전기 생체 임피던스(thoracic electrical bioimpedance; TEB) 또는 전류피부반응 활동(electrodermal activity; EDA)과 같은 생리학적 측정치들을 기록

Table 4: Classification accuracies for example 2

RPEKNN (KNN: $0.622_{0.045}$)					RPEANN (ANN: $0.668_{0.050}$)				
m	d				m	d			
	2	5	10	20		2	5	10	20
1	$0.622_{0.064}$	$0.657_{0.044}$	$0.652_{0.038}$	$0.642_{0.045}$	1	$0.657_{0.064}$	$0.665_{0.045}$	$0.664_{0.042}$	$0.654_{0.048}$
10	$0.707_{0.400}$	$0.701_{0.043}$	$0.684_{0.043}$	$0.652_{0.046}$	10	$0.721_{0.042}$	$0.718_{0.043}$	$0.695_{0.044}$	$0.660_{0.047}$
30	$0.719_{0.044}$	$0.706_{0.044}$	$0.680_{0.045}$	$0.648_{0.044}$	30	$0.734_{0.043}$	$0.728_{0.045}$	$0.693_{0.044}$	$0.662_{0.044}$
60	$0.722_{0.044}$	$0.710_{0.047}$	$0.680_{0.044}$	$0.657_{0.045}$	60	$0.738_{0.045}$	$0.730_{0.045}$	$0.691_{0.046}$	$0.669_{0.046}$

Table 5: Confusion matrix of KNN, ANN, RPEKNN, and RPEANN for example 2

		KNN			ANN			RPEKNN			RPEANN		
		c_1	c_2	c_3	c_1	c_2	c_3	c_1	c_2	c_3	c_1	c_2	c_3
	c_1	30.53	0.00	2.87	30.59	0.00	2.81	30.46	0.00	2.94	30.37	0.00	3.03
Actual	c_2	4.62	14.26	14.23	3.11	15.59	14.41	2.92	20.93	9.26	2.79	20.95	9.37
	c_3	1.71	14.53	17.25	1.72	12.81	18.96	2.35	11.66	19.48	2.09	10.63	20.77

하고, 이를 통해 감정적 활동, 정신적 활동, 육체적 활동, 그리고 중립적 활동의 네 가지 다른 활동을 인식하는 문제를 연구하였다. 자료는 20세 이상 49세 이하의 28명의 남성과 12명의 여성으로부터 측정 및 수집되었다. 각 참가자들은 감정적 활동(영화 감상), 정신적 활동(게임), 육체적 활동(계단 오르내리기), 그리고 중립적 활동(다큐멘터리 감상)에 대해 각각 28번의 기록들을 측정하였으며, 따라서, 총 패턴 수는 4480개였고, 각 집단은 1120개의 다른 패턴으로 구성되어 있다. 설명변수로서 ECG 신호로부터 174개의 속성들이, TEB 신호로부터 151개의 속성들이, 그리고 팔과 손에서 측정된 EDA 신호로부터 각각 104개의 속성들이 측정되어, 총 533개의 변수들이 사용되었다. 본 논문의 실제자료 분석에서는 결측이 포함된 변수 4개(ECG_p_VFL_skewness, ECG_p_LF_skewness, IT_VLF_skewness, IT_LF_skewness.1)를 제외한 총 529개의 설명변수들을 이용하였으며, 표준화된 설명변수들을 이용하였다. 모의실험에서와 마찬가지로 KNN 및 RPEKNN에서의 이웃의 개수에 대한 모수인 k는 1부터 30까지를 고려하였고, ANN 및 제안방법인 RPEKNN에서의 제어계수 q는 1부터 2까지 0.03씩 등간격으로 고려하였으며, 이들 모수의 선택은 교차타당성(leave-one-out cross validation) 방법을 통해 이루어졌다. 분석을 위해 훈련자료와 시험자료를 각각 3,480, 1,000개로 나누어 실험하였으며, 이와 같은 시행을 독립적으로 20번 반복하였다.

Table 6은 행동인식 자료의 분석 결과를 나타낸다. KNN과 ANN 모두 0.93이 넘는 높은 정분류율을 보였다. 모의실험에서와는 달리 d가 작을 때의 랜덤 투영 분류 방법($m = 1$)은 원래 자료의 거리를 충분히 반영하지 못하여 낮은 정분류율을 보였으나, d가 증가함에 따라 점차 각각 KNN과 ANN의 정분류율에 근접하는 경향을 확인할 수 있었다. $d = 10$인 경우, m이 작을 때의 랜덤 투영 앙상블 기법들은 그들의 기저 분류기인 KNN과 ANN의 정분류율보다 낮은 성능을 보였으나, $m = 60$인 경우, 기저 분류기들의 성능에 근접하였다. RPEKNN과 RPEANN 모두 $d = 20$이상에서, $m = 30$이상에서 안정적인 정분류율을 보였으며, 제안 방법인 RPEANN이 기존의 RPEKNN보다 평균적으로 더 높은 정분류율을 보임을 확인할 수 있다.

Table 7은 행동인식 자료의 분석 결과에 대한 KNN과 ANN 그리고 $m = 30$, $d = 20$에서의 RPEKNN과 RPEANN의 혼동행렬(confusion matrix)을 나타내고 있으며, 각 성분은 20번 독립 반복 시행에서의 각 집단별 분류 결과의 평균값이다. 여기서 c_1, c_2, c_3, 그리고 c_4는 각각 중립적 활동, 감정적 활동, 정신적 활동, 육체적 활동을 나타낸다. 모든 분석 방법에서 육체적 활동은 다른 활동들과 확연히 구분되게 분류하였으며 정분류율은 0.99에 달했다. 중립적 활동 또한 육체적 활동과 확연히 구분되었으며, 감정적 활동 및 정신적 활동과도

Table 6: Classification accuracies for activity recognition data

m	RPEKNN (KNN: $0.930_{0.010}$)				m	RPEANN (ANN: $0.935_{0.009}$)			
	d					d			
	5	10	20	40		5	10	20	40
1	$0.662_{0.033}$	$0.835_{0.022}$	$0.898_{0.014}$	$0.917_{0.012}$	1	$0.663_{0.041}$	$0.836_{0.057}$	$0.912_{0.038}$	$0.923_{0.013}$
10	$0.833_{0.014}$	$0.917_{0.011}$	$0.932_{0.010}$	$0.936_{0.009}$	10	$0.842_{0.018}$	$0.921_{0.009}$	$0.939_{0.009}$	$0.938_{0.010}$
30	$0.880_{0.022}$	$0.929_{0.009}$	$0.937_{0.010}$	$0.936_{0.010}$	30	$0.883_{0.016}$	$0.933_{0.010}$	$0.940_{0.009}$	$0.941_{0.010}$
60	$0.893_{0.016}$	$0.933_{0.009}$	$0.938_{0.010}$	$0.937_{0.008}$	60	$0.894_{0.013}$	$0.936_{0.010}$	$0.940_{0.010}$	$0.941_{0.010}$

Table 7: Confusion matrix of KNN, ANN, RPEKNN, and RPEANN for activity recognition data

		KNN				ANN			
		c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
Actual	c_1	237.05	5.40	8.55	0.15	238.55	5.45	8.00	0.15
	c_2	2.15	219.55	24.35	0.45	1.50	223.70	20.95	0.35
	c_3	5.45	21.30	224.75	0.80	4.40	21.60	225.55	0.75
	c_4	0.25	0.75	1.30	247.75	0.25	0.50	1.55	247.75
		RPEKNN				RPEANN			
		c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
Actual	c_1	238.55	5.55	6.85	0.20	238.55	5.00	7.40	0.20
	c_2	1.45	223.70	20.85	0.50	1.15	226.00	18.85	0.50
	c_3	3.65	21.05	226.55	1.05	3.95	20.05	227.20	1.10
	c_4	0.30	0.40	1.35	248.0	0.25	0.40	1.25	248.15

c_1: neutral activity; c_2: emotional activity; c_3: mental activity; c_4: physical activity.

잘 구분되게 분류하여 0.95에 가까운 정분류율을 보였다. 감정적 활동과 정신적 활동은 다른 두 활동에 비해 상대적으로 분류가 잘 되지 않았으나, 제안 방법인 RPEANN이 다른 방법들에 비해 정분류율의 측면에서 더 성능이 뛰어났다.

5. 결론

본 연구에서는 고차원 자료에서의 판별분류분석을 위해 ANN과 랜덤 투영 앙상블 기법을 결합한 새로운 방법을 제안하였다. 본 연구에서 제안한 RPEANN은 자료의 국소적 특징을 반영하여 이웃을 선택함을 통해 KNN보다 더 높은 정분류율 갖는 ANN을 기저 분류기로 사용하여 여러 랜덤 투영 분류기들을 생성하고, 이들의 결과를 종합함으로써 기존의 RPEKNN보다도 더 높은 정분류율을 가짐을 모의실험 및 실제 자료 분석을 통해 확인할 수 있었다. 제안 방법은 알고리즘의 특성상 쉽게 병렬 계산이 가능하므로, 다중 프로세서를 갖춘 병렬 연산이 가능한 연산 환경에서 보다 효율적으로 수행될 수 있다. 제안 방법은 랜덤 투영 분류기의 수 m이 어느 정도 이상의 값으로 설정된다면, 제안 방법인 RPEANN의 성능에 크게 영향을 주지 않음을 확인하였다. 반면 d의 선택의 경우 모의실험에서는 $d = 2$에서 최적의 성능을 보인 반면, 실제 자료의 분석에서는 $d = 20$, $d = 40$에서 최적의 성능을 보였다. KNN과 ANN에서 이웃의 수를 결정하는 k나 q를 선택할 때와 마찬가지로 훈련 과정에서 교차타당법을 통해 최적의 d를 구할 수 있겠으나, 이와 같은 방법은 너무 많은 훈련 시간을 요구할 수 있다. 충분 차원 축약(sufficient dimension reduction)과 같은 기법도 한가지 방법이 될 수 있겠으나, 결정 경계(decision boundary)가 비선형일 경우 활용이 제한될 수 있다. 보다 효율적인 투영 차원 d를 구하는 방법에 관한 추후 연구를 통해 제안 방법의 활용성이 증대될 수 있을 것이다.

References

Bingham E and Mannila H (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 245–250.

Cannings TI and Samworth RJ (2017). Random-projection ensemble classification, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **79**, 959–1035.

Chang DJ and Kwon Y M (2008). Medical diagnosis inference using neural network and discriminant analyses, *Journal of Korean Data & Information Science Society*, **19**, 511–518.

Devroye L, Györfi L, and Lugosi G (1996). *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

Fix E and Hodges JL (1951). Discriminatory analysis - nonparametric discrimination: consistency properties, *Technical Report 4*, Texas, United States.

Friedman J (1994). *Flexible Metric Nearest Neighbor Classification*, Standford University, United States.

Hall P, Park BU, and Samworth RJ (2008). Choice of neighbour order in nearest-neighbour classification, *Annals of Statististics*, **36**, 2135–2152.

Hastie T and Tibshirani R (1996). Discriminant adaptive nearest neighbor classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 607–616.

Jhun MS and Choi IK (2009). Adaptive nearest neighbors for classification. *The Korean Journal of Applied Statistics*, **22**, 479–488.

Johnson WB and Lindenstrauss J (1984). Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics*, **26** 189–206.

Kang JK and Jhun MS (2020a). Divide-and-conquer random sketched kernel ridge regression for large-scale data analysis, *Journal of Korean Data & Information Science Society*, **31**, 15–23.

Kang JK and Jhun MS (2020b). Variable selection in reproducing kernel Hilbert space using random sketch method, *Journal of Korean Data & Information Science Society*, **31**, 501–511.

Kang SA, Kim YS, and Choi SH (2015). Study on the social issue sentiment classification using text mining, *Journal of Korean Data & Information Science Society*, **26**, 1167–1173.

Kim TH and Kim YH (2013). A study on the analysis of customer loan for the credit finance company using classification model, *Journal of Korean Data & Information Science Society*, **24**, 411–425.

Mohino-Herranz I, Gil-Pita R, Rosa-Zurera M, and Seoane F (2019). Activity recognition using wearable physiological measurements: selection of features from a comprehensive literature study, *Sensors*, **19**, 5524.

Shasha DE and Zhu Y (2004). *High Performance Discovery in Time Series: Techniques and Case Studies*. Springer Science & Business Media, New York.

Trefethen LN and Bau D III. (1997). Numerical linear algebra, *Society for Industrial and Applied Mathematics*, Philadelphia, United States.

Received March 30, 2021; Revised April 25, 2021; Accepted April 25, 2021

랜덤 투영 앙상블 기법을 활용한 적응 최근접 이웃 판별분류기법

강종경[a], 전명식[1,b]

[a]고려대학교 통계학과 [b]한국뉴욕주립대학교 응용수학통계학과

요 약

판별분류분석에서 널리 이용되는 k-최근접 이웃 분류 방법은 고정된 이웃의 수만을 고려하여 자료의 국소적 특징을 반영하지 못하는 한계가 있다. 이에 자료의 국소적 구조를 고려하여 이웃의 개수를 선택하는 적응 최근접이웃방법이 개발된 바 있다. 고차원 자료의 분석에 있어서는 k-최근접 이웃 분류를 사용하기 전에 랜덤 투영 기법 등을 활용하여 차원 축소를 수행하는 것이 일반적이다. 이렇게 랜덤 투영시킨 다수의 분류 결과들을 면밀히 조합하여 투표를 통해 최종 할당을 하는 기법이 최근 개발된 바 있다. 본 연구에서는 고차원 자료에서의 분석을 위해 적응 최근접이웃방법과 랜덤 투영 앙상블 기법을 조합한 새로운 판별분류 기법을 제안하였다. 제안된 방법은 기존에 개발된 방법에 비해 분류 정확성 측면에서 더 뛰어남을 모의실험 및 실제 사례 분석을 통해 확인하였다.

주요용어: 고차원 자료, 랜덤 투영, 적응 이웃 분류, 최근접 이웃 분류, 판별분류분석

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2020R1F1A1A01061746).

[1]교신저자 : 교수, 한국뉴욕주립대학교 응용수학통계학과.