커널 기반의 퍼지 *K*-Nearest Neighbor 알고리즘

Fuzzy *K*-Nearest Neighbor Algorithm based on Kernel Method

최병인, 이정훈
한양대학교 전자전기제어계측공학과

Byung-In Choi and Frank Chung-Hoon Rhee
Computational Vision and Fuzzy Systems Laboratory
Department of Electronic Engineering, Hanyang University, Ansan, Korea
Email: {kschoi, frhee}@fuzzy.hanyang.ac.kr

ABSTRACT

커널 함수는 데이터를 high dimension 상의 속성 공간으로 mapping함으로써 복잡한 분포를 가지는 데이터에 대하여 기존의 선형 분류 알고리즘들의 성능을 향상시킬 수 있다. 본 논문에서는 기존의 유클리디안 거리측정방법 대신에 커널 함수에 의한 속성 공간의 거리측정방법을 fuzzy *K*-nearest neighbor 알고리즘에 적용한 fuzzy kernel *K*-nearest neighbor(FKKNN) 알고리즘을 제안한다. 제시한 알고리즘은 데이터에 대한 적절한 커널 함수의 선택으로 기존 알고리즘의 성능을 향상 시킬 수 있다. 제시한 알고리즘의 타당성을 보이기 위하여 여러 데이터 집합에 대한 실험결과를 분석한다.

Key words : kernel method, fuzzy *K*-nearest neighbor, kernel function, nonlinear classification

1. 서 론

기존의 *K*-nearest neighbor(*K*-NN)은 주어진 패턴으로부터 *K*개의 가장 가까운 패턴들의 주된 클래스로 패턴을 분류하는 간단하고, nonparametric한 알고리즘이다[1]. 반면에, KNN은 각 sample pattern들이 input data의 class label을 결정하는데 같은 기여도를 가지는 문제점이 있다. 이는 다른 클래스의 패턴들이 겹쳐있는 데이터에 대하여 오 분류를 야기할 수 있다. 이를 개선하기 위하여 fuzzy set 이론을 *K*-NN rule에 적용시킨 fuzzy *K*-NN 알고리즘이 제안되었다[2]. fuzzy *K*-NN은 입력 패턴에 가까운 *K*개의 패턴들에 따른 fuzzy membership 값을 sample data로부터 얻어진 fuzzy class membership을 사용하여 기여도에 따라 할당한다. 따라서, 서로 겹친 부분에서 *K*-NN 보다 향상된 분류 결과를 얻을 수 있다. 그러나 기존의 *K*-NN 또는 fuzzy *K*-NN은 유클리디안 공간상의 패턴 간 거리를 사용한다. 이러한 거리 측정 방법을 대신하여 커널 함수에 의한 높은 차원의 속성 공간의 거리 측정을 사용한 kernel *K*-nearest neighbor (KKNN) 알고리즘이 제안되었다[3]. 제안된 알고리즘은 복잡한 분포를 가지는 데이터에 대하여 향상된 분류성능을 가질 수 있었다. 본 논문에서는 이러한 커널 함수를 기존의 fuzzy *K*-NN의 fuzzy class membership과 fuzzy classification rule에 적용한 fuzzy kernel *K*-nearest neighbor 알고리즘을 제안한다. 제안한 알고리즘은 커널 함수의 적용으로, arbitrary한 분포를 갖는 sample data에 대하여 fuzzy *K*-NN 보다 향상된 classification 성능을 얻을 수 있다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 2 절에서는 커널 함수에 대하여 간략히 언급하고, 3 절에서는 제안한 fuzzy kernel *K*-NN 알고리즘에 대하여 설명한다. 다음으로 제안한 알고리즘의 타당성을 보이기 위하여 여러 데이터들에 대한 결과를 fuzzy *K*-NN과

비교하고, 마지막으로 결론을 맺을 것이다.

2. 커널 함수

커널 함수의 적용은 입력 데이터들을 높은 차원의 속성 공간으로 변환하여 처리를 하는 것이다[4][5][6]. 이러한 변환은 그림 1 에서 볼 수 있듯이 입력 데이터 공간에서 분류가 어려운 데이터들을 속성 공간상으로 변환하여 분류가 용이하게 만든다.

n 차원 데이터 공간상의 패턴 x 가 N 차원 속성 공간으로 feature mapping 할 경우를 가정하면 다음과 같다.

$$\mathbf{x}=(x_1,\cdots,x_n)\rightarrow\phi(\mathbf{x})=(\phi_1(\mathbf{x}),\cdots\phi_N(\mathbf{x})) \quad (1)$$

여기서, $\phi(\cdot)$는 비선형적으로 데이터 공간의 패턴을 높은 차원의 속성 공간으로 변환시켜주는 변환 함수이다.

커널 함수는 다음과 같이 비선형 변환 함수에 의한 속성 공간상에서의 내적으로 정의된다.

$$K(\mathbf{x},\mathbf{y})=\phi(\mathbf{x})\cdot\phi(\mathbf{y}) \quad (2)$$

그러므로, 커널 함수를 통하여 큰 계산 량이 요구되는 속성 공간상으로의 변환 없이 내적을 구할 수 있다.

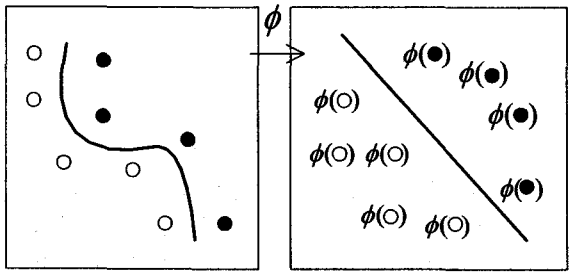

그림 1. 커널 함수를 사용한 속성 공간 변환
(a) 입력 공간 (2) 변환된 속성 공간

다음의 세 가지의 커널 함수가 일반적으로 사용된다.

(1) Polynomial kernel

$$K(\mathbf{x},\mathbf{y})=(1+\mathbf{x}\cdot\mathbf{y})^p \quad (3)$$

(2) Radial basis kernel

$$K(\mathbf{x},\mathbf{y})=\exp\left\{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right\} \quad (4)$$

(3) Sigmoid kernel

$$K(\mathbf{x},\mathbf{y})=\tanh(\alpha(\mathbf{x}\cdot\mathbf{y})+\beta) \quad (5)$$

여기서 p,σ,α,β는 커널 함수의 파라 미터들이다.

3. Fuzzy Kernel *K*-Nearest Neighbor 알고리즘

본 논문에서 제안하는 fuzzy kernel K-nearest neighbor 알고리즘의 주된 목적은 높은 차원의 속성 공간상에서 거리 측정을 통하여 fuzzy K-nearest neighbor 알고리즘의 수행 성능을 향상시키는 것이다. 그러므로 먼저 속성 공간에서 데이터간 거리 측정 방법을 설명한다.

입력 패턴 **x**, **y**의 속성 공간 상의 거리는 다음과 같이 나타난다.

$$\begin{aligned} d_\phi(\mathbf{x},\mathbf{y}) &= \|\phi(\mathbf{x})-\phi(\mathbf{y})\|^2 \\ &= \phi(\mathbf{x})\phi(\mathbf{y})-2\phi(\mathbf{x})\phi(\mathbf{y})+\phi(\mathbf{x})\phi(\mathbf{y}) \end{aligned} \quad (6)$$

식 (2)의 커널 함수의 정의로부터 식 (6)의 속성 공간상의 내적은 다음과 같이 커널 함수로 대치될 수 있다.

$$d_\phi(\mathbf{x},\mathbf{y})=K(\mathbf{x},\mathbf{y})-2K(\mathbf{x},\mathbf{y})+K(\mathbf{x},\mathbf{y}) \quad (7)$$

제안한 알고리즘은 위에서 정의한 커널 거리 측정 방법을 기존의 fuzzy K-NN 알고리즘에 적용한다. fuzzy kernel K-NN 알고리즘은 fuzzy K-NN에서와 마찬가지로 입력 패턴으로부터 labeled된 sample 패턴들 중에서 가장 가까운 K개의 패턴들을 사용하여 각 클래스에 대한 fuzzy membership 값을 할당한다. 즉, 선택된 K개의 sample 패턴들이 입력 패턴의 각 클래스에 대한 fuzzy membership에 얼마나 기여하는 지를 나타내는 membership 값과 위에서 제시한 커널 함수를 이용한 속성 공간상의 거리 측정 함수를 사용하여 다음과 같은 fuzzy membership을 할당하게 된다.

$$u_i(\mathbf{x})=\frac{\sum_{j=1}^{K}u_{ij}\left(1/d_\phi(\mathbf{x},\mathbf{x}_j)^{2/(m-1)}\right)}{\sum_{j=1}^{K}1/d_\phi(\mathbf{x},\mathbf{x}_j)^{2/(m-1)}} \quad (8)$$

여기서 u_{ij}는 i번째 클래스에 대한 입력패턴과 가장 가까운 K개의 labeled된 sample 패턴들 중 j번째 패턴의 fuzzy membership을 나타낸다. 식 (8)에서 볼 수 있듯이 fuzzy

membership은 K개의 선택된 sample 패턴의 fuzzy class membership과 거리의 역수에 좌우 된다. 그러므로 패턴들 간의 거리가 가까우면 가까울수록 거리의 역수에 의하여 fuzzy membership에 더욱 크게 기여할 것이다. Labeled된 sample 패턴들의 초기 fuzzy membership 값인 u_{ij}는 [2]에서 다음과 같은 합리적인 방법을 제시하고 있다.

$$u_j(x) = \begin{cases} 0.51 + (n_j / K) \cdot 0.49, & if\ j = i \\ (n_j / K) \cdot 0.49, & if\ j \neq i \end{cases} \quad (9)$$

여기서 n_j는 해당 sample 패턴과 가장 가까운 K개의 sample 패턴들 중에서 j번째 클래스에 소속되어 있는 패턴들의 개수를 나타낸다. 거리 측정 방법은 본 논문에서 제시한 커널 거리 측정 방법을 사용하여 수행한다.

4. 실험 결과 및 분석

제안한 알고리즘의 타당성을 보이기 위하여 "T-shape", "Twoclass", "Pima-Indian"의 3개의 데이터 대하여 제안된 알고리즘과 fuzzy K-NN 알고리즘을 수행하고, 그 결과를 비교 분석한다.

먼저 Labeled sample 패턴들의 초기 fuzzy membership 값을 구하기 위하여 초기 K={1, 3, 5, 7, 9}를 사용하였다. 또한 각 입력 패턴의 fuzzy kernel K-NN을 수행하기 위하여 K={1, 2, 3, 4, 5, 6, 7, 8, 9}를 사용하였다. 즉, 각 K에 대하여 5개의 초기 K를 적용하여 5번의 실험을 하였다. 5번의 실험에 의해 나타난 오분류 패턴 개수를 평균하여 각 K 값의 결과로서 나타낸다.

각 K값의 조합에 따른 실험은 데이터 집합의 모든 패턴에 대하여 하나의 데이터를 제외한 나머지를 Labeled sample 데이터 집합으로 설정하고, 설정된 데이터 집합을 사용하여 제외된 패턴을 분류하게 된다. 모든 실험에서 polynomial kernel, $K(\mathbf{x},\mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^p$ 을 커널 함수로 선택하였다. 또한 Polynomial kernel의 파라 미터 p는 0.1에서 10까지 변화하면서 선택하여 제안한 알고리즘을 수행하고, 수행 결과 중 가장 낮은 오분류 결과를 제안한 알고리즘의 결과로 취한다.

각 데이터들의 속성들은 다른 값의 범위를 가지므로 알고리즘을 수행하기 전에 데이터의 속성 값들을 0과 1사이로 normalization한다.

4.1 "T-shape" 데이터

"T-shape"데이터는 447개의 패턴들로 이루어져 있고, 2개의 속성 값을 가지며 2개의 클래스를 가지며, 각 클래스는 228개와 219개로 구성되어 있다.

데이터의 분포는 다음과 같다.

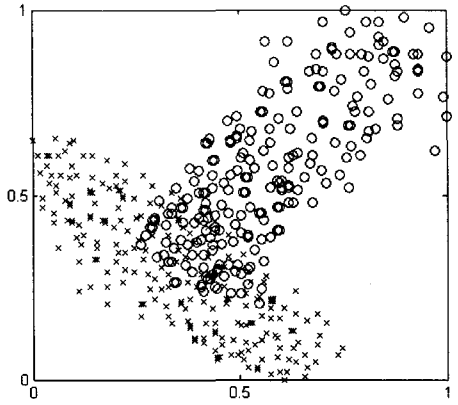

그림 2. "T-shape" 데이터의 scatter plot

제안된 알고리즘과 fuzzy K-NN의 결과는 다음과 같다.

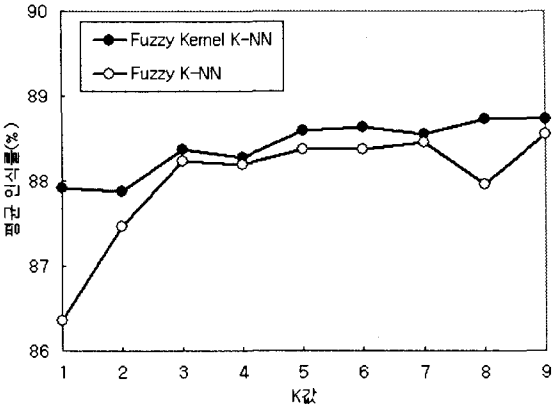

그림 3. "T-shape" 데이터의 평균 오분류 개수

그림 3에서 볼 수 있듯이 주어진 데이터 분포에 대하여 제안한 알고리즘이 fuzzy K-NN보다 향상된 수행 결과를 나타낸다. 전체 평균 분류 결과는 제안된 알고리즘이 88.4%이고, fuzzy K-NN이 88%를 나타내었다. 그러므로 제안한 알고리즘이 약 0.4%의 성능 향상을 얻을 수 있었다.

4.2 "Twoclass" 데이터

"Twoclass" 데이터는 패턴이 242개이고 4개의 속성값을 가지며 2개의 클래스를 갖는 데이터 집합이다. 각 클래스는 121개의 패턴으로 구성되어있다. 제안된 알고리즘과 fuzzy K-NN 알고리즘의 결과는 다음과 같다.

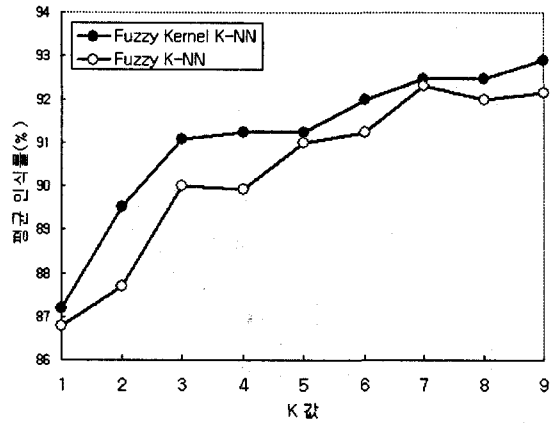

그림 4. "twoclass" 데이터의 평균 오분류 개수

그림 4에서 볼 수 있듯이 모든 K에 대하여 제안한 fuzzy kernel K-NN이 fuzzy K-NN보다 높은 평균 인식률을 나타낸다. 전체 평균 분류 결과는 fuzzy kernel K-NN이 91.2%를 나타내고, fuzzy K-NN이 90.3%를 나타내었다. 적절한 커널 파라 미터 선택으로 4.2의 실험보다 더 높은 약 0.9%의 성능 향상을 얻을 수 있었다.

4.3 "Pima-Indian" 데이터

이번 실험에서는 많은 속성 값을 가지는 데이터에 대한 결과를 분석한다. "Pima-Indian" 데이터는 768개의 패턴으로 이루어져 있고, 8개의 속성 값을 가지며 2개의 클래스로 구성되어 있다. 각 클래스는 500개와 268개의 패턴으로 구성되어 있다. 그림 5는 제안한 알고리즘과 fuzzy K-NN의 결과를 나타낸다.

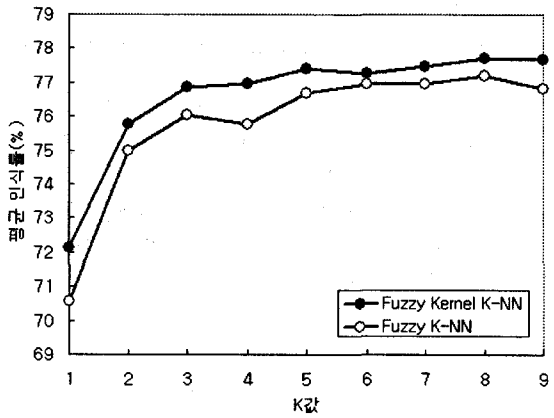

그림 5. "Pima-Indian" 데이터의 평균 오분류 개수

그림에서 볼 수 있듯이 제안한 알고리즘의 성능이 fuzzy K-NN보다 훨씬 나은 것을 볼 수 있다. 또한 전체 평균 분류 결과는 fuzzy kernel K-NN이 76.6%를 나타내고, fuzzy K-NN이 75.8%를 나타내었다. 약 0.8%의 성능 향상을 얻을 수 있었다. 본 실험을 통하여 적절한 커널 함수의 선택과 커널 파라 미터의 선택으로 기존의 fuzzy K-NN보다 상당한 성능의 향상을 얻을 수 있었다. 이러한 성능 향상은 커널 함수에 의하여 데이터의 분류에 더 적당한 속성 공간으로의 변환에 기인한다고 볼 수 있다.

5. 결 론

본 논문에서는 기존의 fuzzy K-NN에 커널 함수에 의한 거리 측정 방법을 적용 시킨 fuzzy kernel K-NN 알고리즘을 제안하였다. 제안한 알고리즘은 커널 함수를 통하여 유클리디안 공간이 아닌 데이터 분류에 적합한 속성 공간으로 데이터의 변환 없이 데이터 간 거리를 측정 할 수 있었다. 그러므로 복잡한 분포의 데이터에 대하여 데이터의 적절한 속성 공간으로의 변환함으로써 기존 알고리즘의 성능을 상당히 향상시킬 수 있었다. 또한 여러 데이터의 실험 결과로부터 제안한 알고리즘의 타당성을 제시하였다. 반면에, 적절한 커널 함수의 선택과 최적의 커널 함수의 파라 미터 선택 방법의 제시가 향후 과제로 남아있다.

감사의 글 : 본 연구는 한국과학기술원 영상정보특화연구센터를 통한 국방과학연구소의 연구비 지원으로 수행되었으며 연구비 지원에 감사 드립니다.

IV. 참고문헌

[1] J. Tou and R. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, 1974.

[2] J. Keller, M. Gray, and J. Givens, JR, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 4, pp.258-263, August 1985.

[3] K. Yu, L. Ji and X. Zhang, "Kernel nearest-neighbor algorithm", *Neural Processing Letters*, vol.15, no. 2, pp.147-156, 2002.

[4] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[5] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[6] B. Schölkolpf, C. Burges and A. Smola, Advances in Kernel Methods, MIT Press, 1998.