




A new general nearest neighbor classification based on the mutual neighborhood information

저널 : Knowledge-Based System 121, 17 January 2017, Pages 142-152
저자 : Zhibin Pan, Yidi Wang, Weiping Ku

18510049 오순묵
18512084 이도현
18512112 한동기



목차

1. 논문 선택 이유
2. 기존 방법론
3. 관련 연구
4. 개선 아이디어
5. 실험
6. 결론

1. 논문 선택 이유

● 기존 논문

- 데이터 간의 individual, neighborhood distance를 각각 정의
- 목적함수를 설계하고 최적화하여 두 가지 기준을 결합하는 방법 제안

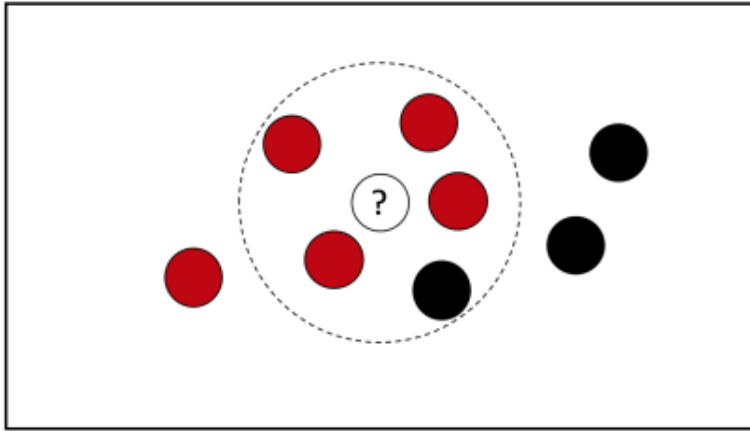
● 본 논문

- training sample, testing sample에 대한 neighborhood를 각각 정의
- 두 가지 기준에 대한 합집합을 사용하는 k GNN 방식 제안



각 논문에서 제시하는 neighborhood에 대한 정의 및 여러 기준을 통합하는 방법의 차이

2. 기존 방법론

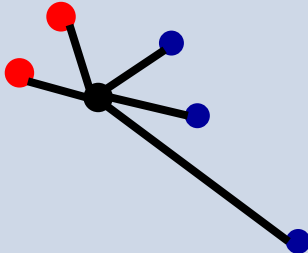
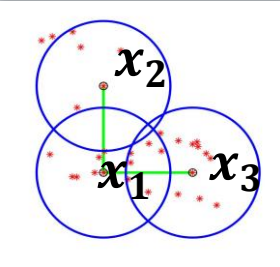


- k 개의 최근접 이웃 데이터를 찾고 다수결 방식으로 분류하는 알고리즘
- k -NN의 성능에 영향을 주는 세 가지 요소
 - 1) value of k
 - 2) distance metric
 - 3) sample size

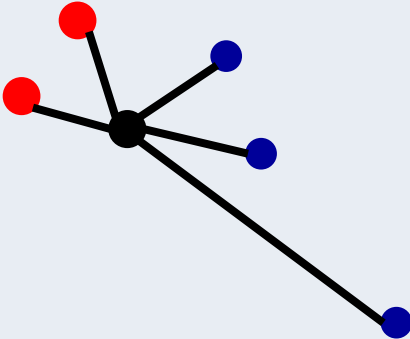
3. 관련 연구

| 연구 주제 | 논문 제목 | 저자 |
|-------------------|--|---|
| Size of k | A proposal for local k values for k -nearest neighbor rule | N. Garcia-Pedrajas, J.A. Del-Castillo, G. Cerruela-Garcia |
| | Neighborhood size selection in the k -nearest neighbor rule using statistical confidence | Wang J., P. Neskovic, L.N. Cooper |
| Distance Function | The distance-weighted k -nearest neighbor rule | S.A. Dudani |
| | IKNN: informative k -nearest neighbor pattern classification | Y. Song, J. Huang, D. Zhou, et al. |
| Outlier | A local mean-based nonparametric classifier | Y. Mitani, Y. Hamamoto |
| | Pseudo nearest neighbor rule for pattern classification | Zeng Y., Yang Y., Zhao L. |

3. 관련 연구

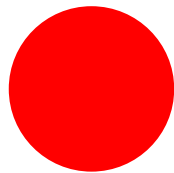
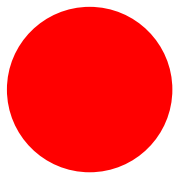
| Distance function | | |
|---|--|---|
| Euclidean | $\ p - q\ = \sqrt{(p - q) \cdot (p - q)} = \sqrt{\ p\ ^2 + \ q\ ^2 - 2p \cdot q}$ | |
| Distance-weighted k -nearest neighbor |  | 가까운 순서대로 weight 추가 |
| FN k NN |  | 일정 Neighborhood(δ) 안에서 Test sample과 Train sample 사이의 샘플 수를 고려 |

3. 관련 연구

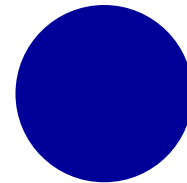
| A local mean-based nonparametric classifier | Outlier ($k = 5$) | Pseudo nearest neighbor rule |
|--|--|---|
| $X^i = \{x_j^i j = 1, \dots, N_i\}$ <i>class i</i> 에 대한 <i>value</i> 집합 j = sample number $y^i = \frac{1}{k} \sum_{j=1}^r x_{kj}^i$ | <div data-bbox="730 529 909 615"> ● Class 1 ● Class 2 </div>  | $w_i = \frac{1}{i}, i = 1, \dots, k$ $y^i = \frac{1}{k} \sum_{j=1}^r x_{kj}^i * w_k$ |
| <ul style="list-style-type: none"> 임의의 NN 개수 설정 (k) distance를 더하고 r로 나눠 local mean vector 생성 (y_i) 각 클래스에 대한 y_i 값 비교 | | <ul style="list-style-type: none"> local mean 과정과 동일 NN 순서별로 weight 설정하여 각 distance에 대한 weighted average를 구함 각 클래스에 대한 y_i 값 비교 |

4. 개선 아이디어

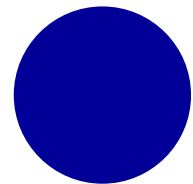
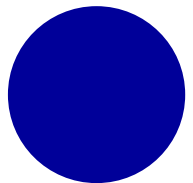
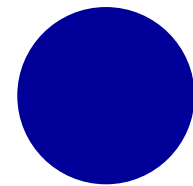
● Class 1
● Class 2



Test sample

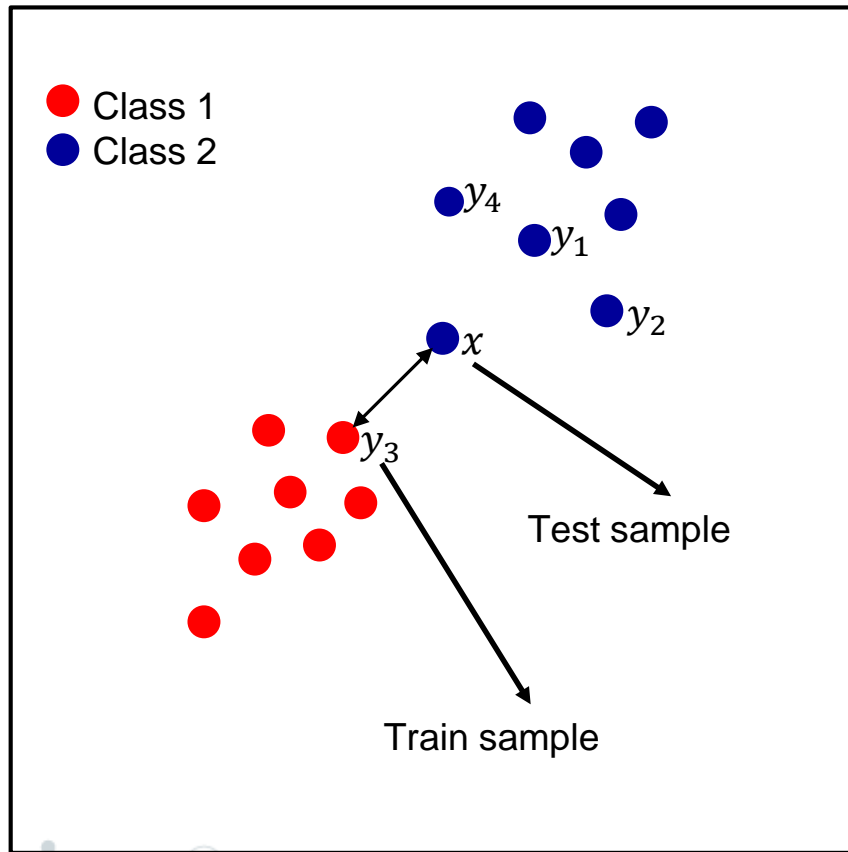


Train sample



Test sample 입장에서만 NN
Unilateral

4. 개선 아이디어



- MRO (Mutual Relationship Observation)

$$MRO(x, y_i) = (ra_{x, y_i}, ra_{y_i, x})$$

$$MRO(x, y_1) = (1, 2)$$

$$MRO(x, y_2) = (2, 2)$$

$$MRO(x, y_3) = (3, 5)$$

$$MRO(x, y_4) = (4, 1)$$



Training sample 측면에서의 NN이 다름을 지적

4. 개선 아이디어

- GNN (General Nearest Neighbor)
- Put a Test sample into Train Samples

$$TS^* = \{y_1, y_2, y_i, \dots, y_N, x \mid y_i \in TS, 1 \leq i \leq N\}$$
$$y_i^* \in TS^*, 1 \leq i \leq N+1, y_i^* \text{ can be either } x \text{ or } y_i \in TS^*$$

The neighborhood Information of y_i^* in the feature space R^P

$$\rightarrow N^{\delta_{y_i^*}}(y_i^*) = \{y \mid \forall y \in R^P \cap d(y_i^*, y) \leq \delta_{y_i^*}\}$$

4. 개선 아이디어

- GNN (General Nearest Neighbor)



- Training Sample의 NN

$$N^{\delta_{y_i}}(\mathbf{y}_i)$$

- Testing Sample의 NN

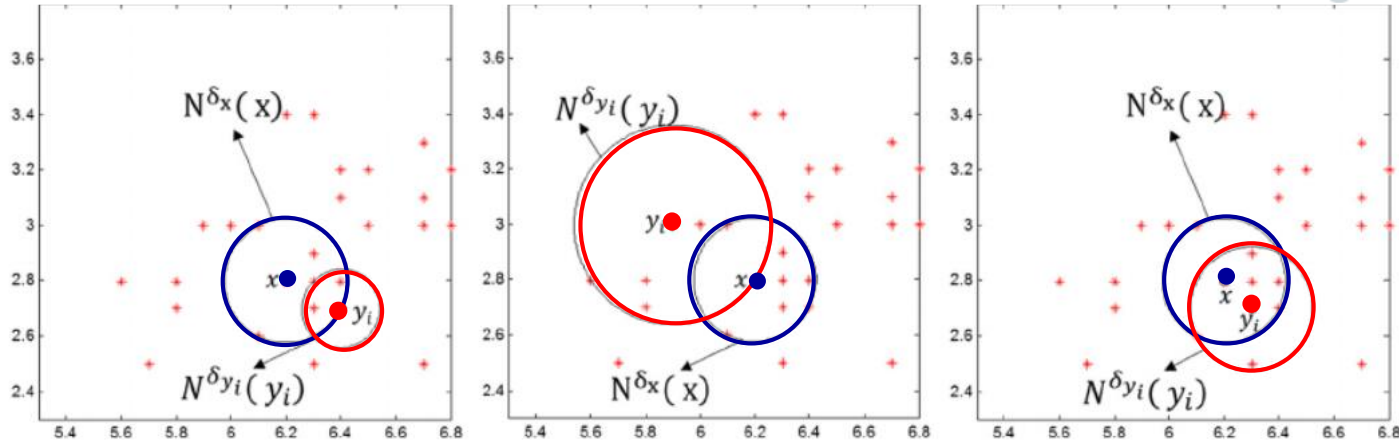
$$N^{\delta_x}(\mathbf{x})$$



GNN Rule : $\left[\mathbf{x} \in N^{\delta_{y_i}}(\mathbf{y}_i) \right] \vee \left[\mathbf{y}_i \in N^{\delta_x}(\mathbf{x}) \right], \quad 1 \leq i \leq N$

4. 개선 아이디어

- GNN (General Nearest Neighbor)



● Train sample
● Test sample

- GNN Rule

$$\left[x \in N^{\delta_{y_i}}(y_i) \right] \vee \left[y_i \in N^{\delta_x}(x) \right], \quad 1 \leq i \leq N$$

- Symmetrical Relation

$$y_i \in GNN(x) \Leftrightarrow x \in GNN(y_i)$$

4. 개선 아이디어

- ***k*GNN Classifier**

1. Train sample 전체와 Test sample 하나의 합집합 U 를 생성한다
2. 합집합 U 의 각 Sample 측면에서의 k NN을 구한다
3. GNN Rule에 기반하여 k GNN을 구한다

$$GNN^k(x) = \left[x \in N^k(y_i) \right] \vee \left[y_i \in N^k(x) \right], \quad 1 \leq i \leq N$$

4. 구한 k GNN에 대해 majority voting 방식으로 Test sample의 클래스를 결정한다

5. 실험

- Real World Datasets

Table 1
Dataset description.

| Dataset | Database | Samples | Attributes | Classes | Training set |
|------------|----------|---------|------------|---------|--------------|
| Australian | KEEL | 690 | 14 | 2 | 230 |
| Balance | UCI | 625 | 4 | 3 | 200 |
| Banana | KEEL | 5300 | 2 | 2 | 1800 |
| Breast | UCI | 277 | 9 | 2 | 90 |
| German | KEEL | 1000 | 20 | 2 | 300 |
| Glass | KEEL | 214 | 9 | 7 | 70 |
| Heart | UCI | 303 | 13 | 2 | 100 |
| Iris | UCI | 150 | 4 | 3 | 50 |
| Landsat | UCI | 2000 | 36 | 6 | 700 |
| Liver | UCI | 345 | 6 | 2 | 115 |
| Monk-2 | KEEL | 432 | 6 | 2 | 144 |
| Newthyroid | KEEL | 215 | 5 | 3 | 70 |
| Optdigits | KEEL | 5620 | 64 | 10 | 1800 |
| Segment | KEEL | 2310 | 19 | 7 | 770 |
| Spambase | KEEL | 4597 | 57 | 2 | 1500 |
| Thyroid | KEEL | 7200 | 21 | 3 | 2400 |
| Wine | UCI | 178 | 13 | 3 | 60 |
| Wpbc | UCI | 198 | 33 | 2 | 66 |
| Vehicle | UCI | 846 | 18 | 4 | 282 |
| Vote | UCI | 435 | 16 | 2 | 145 |

- UCI, KEEL의 데이터 활용
- 다양한 수의 Sample, Feature, Class

5. 실험

- Artificial Datasets

특정 평균과 공분산을 가지는 두 가우시안 분포에서 데이터 생성

- I-I Datasets

$$\mu_1 = 0 \quad \mu_2 = \begin{bmatrix} 2.56 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \Sigma_1 = \Sigma_2 = I_p$$

- Ness Datasets

$$\mu_1 = 0 \quad \mu_2 = \begin{bmatrix} 0.5\Delta \\ 0 \\ \vdots \\ 0 \\ 0.5\Delta \end{bmatrix} \quad \Sigma_1 = I_p \quad \Sigma_2 = \begin{bmatrix} I_{\frac{p}{2}} & O \\ O & I_{\frac{p}{2}} \end{bmatrix}$$

5. 실험

● k GNN과 k NN 비교

Table 2

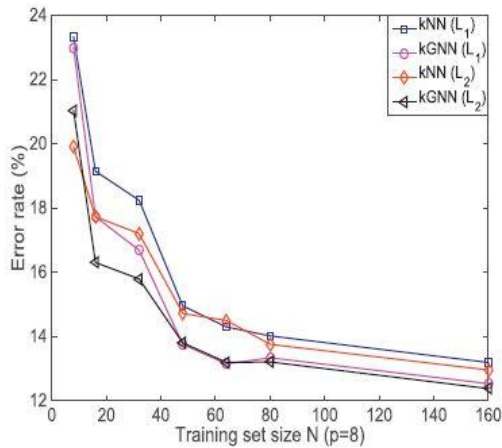
Comparison of k NN and k GNN in terms of the error rate (%) using Euclidean norm (L_2 norm) and Manhattan norm (L_1 norm) when $k = 1, 5$ and 9 .

| Dataset | L_1 | | | | | | L_2 | | | | | |
|------------|---------|--------------|---------|--------------|---------|--------------|--------------|--------------|---------|--------------|---------|--------------|
| | $k = 1$ | | $k = 5$ | | $k = 9$ | | $k = 1$ | | $k = 5$ | | $k = 9$ | |
| | k NN | k GNN | k NN | k GNN | k NN | k GNN | k NN | k GNN | k NN | k GNN | k NN | k GNN |
| Australian | 36.74 | 36.52 | 33.87 | 31.96 | 33.26 | 32.61 | 36.74 | 35.22 | 35.43 | 33.70 | 34.35 | 31.08 |
| Balance | 22.01 | 19.53 | 16.64 | 13.73 | 14.11 | 12.40 | 21.55 | 19.34 | 15.34 | 13.19 | 13.28 | 12.78 |
| Banana | 12.97 | 12.24 | 11.34 | 11.17 | 10.97 | 9.86 | 12.80 | 12.14 | 11.34 | 11.05 | 10.83 | 10.03 |
| Breast | 32.70 | 30.25 | 28.10 | 27.46 | 26.79 | 26.25 | 34.65 | 32.28 | 29.76 | 29.38 | 27.33 | 27.13 |
| German | 33.84 | 33.21 | 29.64 | 29.10 | 28.92 | 28.00 | 35.79 | 35.19 | 31.79 | 31.52 | 31.66 | 30.25 |
| Glass | 31.98 | 31.90 | 35.24 | 34.34 | 38.82 | 38.22 | 35.80 | 34.13 | 37.71 | 35.59 | 39.79 | 39.34 |
| Heart | 23.47 | 22.64 | 19.24 | 19.06 | 19.83 | 19.20 | 24.70 | 23.82 | 21.63 | 21.05 | 22.26 | 21.50 |
| Iris | 6.20 | 6.05 | 5.40 | 5.05 | 5.95 | 5.55 | 6.20 | 6.20 | 5.05 | 4.10 | 6.20 | 5.45 |
| Landsat | 13.64 | 12.92 | 13.21 | 12.35 | 13.29 | 13.00 | 13.57 | 13.48 | 13.80 | 12.14 | 14.69 | 13.78 |
| Liver | 40.70 | 40.39 | 36.17 | 34.82 | 35.89 | 34.91 | 40.50 | 40.26 | 37.02 | 34.97 | 36.15 | 34.95 |
| Monk-2 | 13.39 | 7.08 | 6.08 | 3.91 | 4.91 | 4.09 | 13.32 | 7.22 | 6.20 | 4.32 | 5.59 | 4.60 |
| Newthyroid | 8.38 | 8.21 | 12.21 | 9.93 | 15.34 | 12.37 | 8.97 | 8.65 | 13.10 | 11.10 | 16.66 | 14.06 |
| Optdigits | 2.74 | 2.66 | 2.47 | 2.20 | 2.66 | 2.50 | 2.36 | 2.36 | 2.07 | 1.90 | 2.30 | 2.15 |
| Segment | 4.74 | 4.61 | 6.43 | 6.16 | 7.79 | 7.40 | 6.83 | 6.68 | 9.75 | 9.15 | 11.30 | 10.84 |
| Spambase | 19.57 | 19.37 | 20.60 | 19.30 | 20.72 | 19.98 | 22.76 | 21.08 | 23.42 | 22.28 | 24.32 | 22.79 |
| Thyroid | 7.33 | 7.27 | 6.40 | 6.22 | 6.46 | 6.25 | 7.89 | 7.81 | 6.60 | 6.33 | 6.54 | 6.39 |
| Wine | 24.87 | 24.87 | 25.85 | 25.21 | 26.52 | 25.29 | 30.38 | 30.55 | 30.08 | 29.40 | 30.33 | 29.06 |
| Wpbc | 33.59 | 32.87 | 25.98 | 24.50 | 24.54 | 23.75 | 33.48 | 32.53 | 26.40 | 25.11 | 25.00 | 23.82 |
| Vehicle | 32.94 | 32.64 | 32.70 | 31.00 | 32.76 | 31.99 | 34.24 | 33.97 | 34.02 | 33.09 | 35.13 | 34.18 |
| Vote | 8.74 | 8.18 | 7.91 | 6.72 | 8.63 | 7.17 | 9.77 | 9.43 | 9.03 | 7.65 | 9.79 | 8.27 |
| Average | 20.53 | 19.67 | 18.77 | 17.70 | 18.91 | 18.03 | 21.62 | 20.62 | 19.98 | 18.85 | 20.18 | 19.12 |

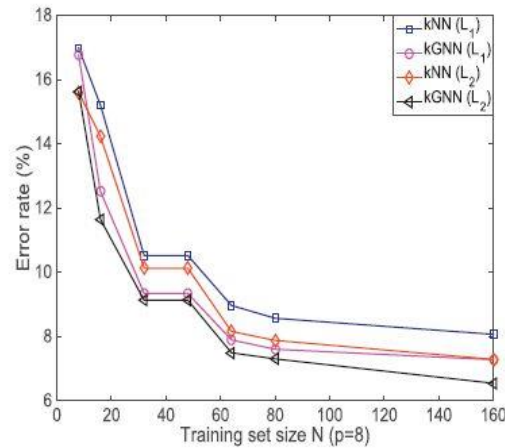
대부분의 Datasets 종류, k 값, 거리 함수(L_1, L_2)에 대해 성능 향상

5. 실험

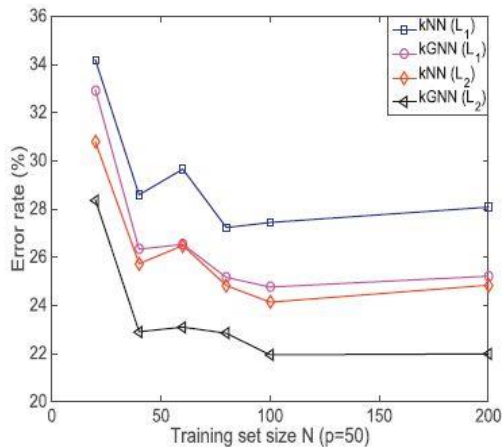
● k GNN과 k NN 비교



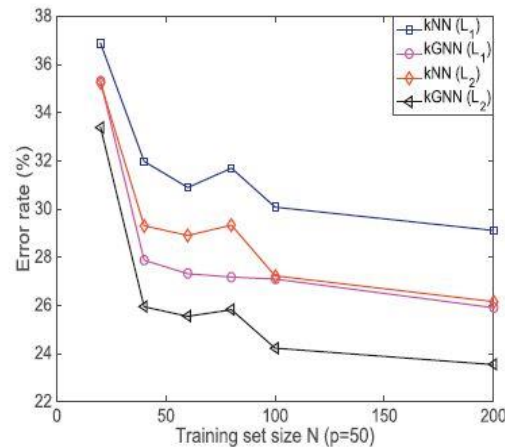
(a) I-I dataset ($p=8$)



(b) Ness dataset ($\Delta=4, p=8$)



(c) I-I dataset ($p=50$)



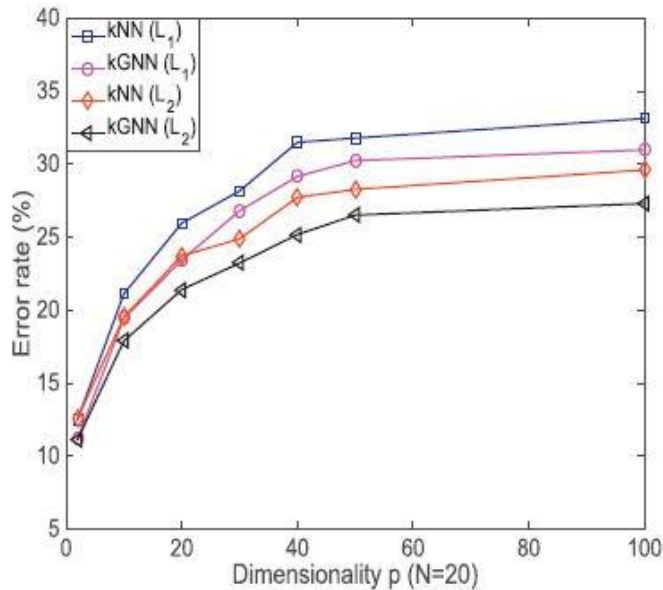
(d) Ness dataset ($\Delta=4, p=50$)

Fig. 4. Influences of the training sample size N on the error rate (%).

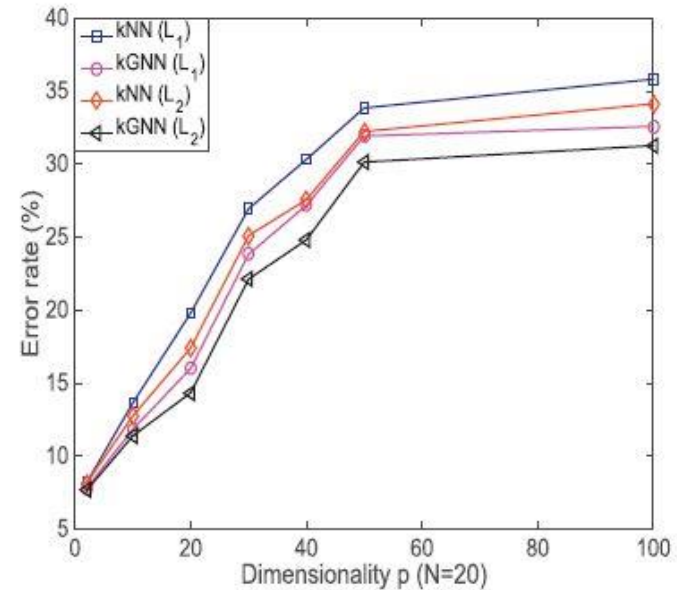
- 다양한 조건에 대해 성능 향상
- N 값을 증가시킬수록 오차율의 차이가 커지는 경향

5. 실험

- k GNN과 k NN 비교



(a) I-I dataset ($N=20$)



(b) Ness dataset ($\Delta=4$, $N=20$)

Fig. 5. Influences of the dimensionality p on the error rate (%).

차원 p 를 증가시킬수록 오차율의 차이가 커지는 경향

5. 실험

● k GNN과 다른 Classifier 비교

Table 3

Comparison of k GNN and other classifiers in terms of the error rate (%) with the optimized value of k .

| Dataset | LMKNN | MKNN | PNN | WKNN | CFKNN | FRNN | HBKNN | k GNN |
|------------|-----------------|-----------------|-----------------|------------|-----------------|-----------------|----------------|-----------------|
| Australian | 32.83(6) | 30.43(2) | 29.78(1) | 35.65(7) | 31.09(3.5) | 35.86(8) | 31.95(5) | 31.09(3.5) |
| Balance | 11.20(1) | 13.02(5) | 12.95(4) | 13.26(6) | 12.02(2) | 13.33(7) | 13.77(8) | 12.69(3) |
| Banana | 9.88(3) | 9.74(2) | 10.20(4) | 12.34(6) | 27.68(8) | 26.71(7) | 11.25(5) | 9.51(1) |
| Breast | 28.18(4) | 28.93(6) | 27.99(3) | 27.54(2) | 29.76(7) | 31.73(8) | 28.20(5) | 27.14(1) |
| German | 30.90(7) | 30.56(5) | 30.75(6) | 30.42(4) | 29.64(2) | 28.58(1) | 31.50(8) | 29.99(3) |
| Glass | 35.80(6) | 33.92(3) | 34.72(5) | 34.38(4) | 36.83(8) | 36.49(7) | 33.49(2) | 32.88(1) |
| Heart | 20.02(2) | 22.19(8) | 20.83(4) | 22.05(7) | 20.88(5) | 19.04(1) | 20.89(6) | 20.69(3) |
| Iris | 4.60(2) | 4.85(3) | 5.05(5) | 4.90(4) | 10.70(8) | 10.10(7) | 5.25(6) | 4.10(1) |
| Landsat | 12.18(4) | 11.94(1) | 12.74(5) | 12.87(6) | 17.50(8) | 16.34(7) | 12.15(3) | 12.14(2) |
| Liver | 34.79(3) | 35.87(7) | 35.28(5) | 35.58(6) | 34.22(2) | 41.15(8) | 35.23(4) | 34.19(1) |
| Monk-2 | 4.62(4) | 4.65(5) | 4.60(3) | 6.09(6) | 8.35(8) | 7.21(7) | 2.88(1) | 4.32(2) |
| Newthyroid | 8.69(2.5) | 8.96(5.5) | 8.96(5.5) | 8.79(4) | 12.58(7) | 8.69(2.5) | 13.58(8) | 8.65(1) |
| Optdigits | 1.49(1) | 1.78(2) | 1.89(5) | 1.86(3) | 1.96(6) | 2.95(8) | 2.48(7) | 1.88(4) |
| Segment | 6.59(4.5) | 6.59(4.5) | 6.72(6) | 6.83(7) | 11.62(8) | 6.33(3) | 5.93(1) | 6.12(2) |
| Spambase | 20.72(2) | 22.35(8) | 21.13(3) | 21.74(7) | 21.36(5) | 21.38(6) | 21.25(4) | 20.67(1) |
| Thyroid | 6.23(3) | 6.18(2) | 6.42(5) | 6.37(4) | 8.33(8) | 7.39(7) | 6.77(6) | 6.17(1) |
| Wine | 28.43(4) | 28.43(4) | 28.81(6) | 29.61(7) | 32.07(8) | 14.49(1) | 28.43(4) | 27.92(2) |
| Wpbc | 23.39(4) | 25.75(7) | 23.26(2) | 23.58(6) | 22.00(1) | 27.19(8) | 23.40(5) | 23.33(3) |
| Vehicle | 30.72(2) | 32.61(5) | 33.07(6.5) | 33.07(6.5) | 29.57(1) | 34.45(8) | 31.80(3) | 32.59(4) |
| Vote | 7.22(1) | 8.38(6) | 9.00(8) | 8.93(7) | 7.60(3) | 8.01(4) | 8.03(5) | 7.44(2) |
| Ranking* | 3.30 | 4.55 | 4.60 | 5.48 | 5.43 | 5.78 | 4.80 | 2.08 |
| Average | 17.92 | 18.36 | 18.20 | 18.79 | 20.28 | 19.87 | 18.41 | 17.66 |

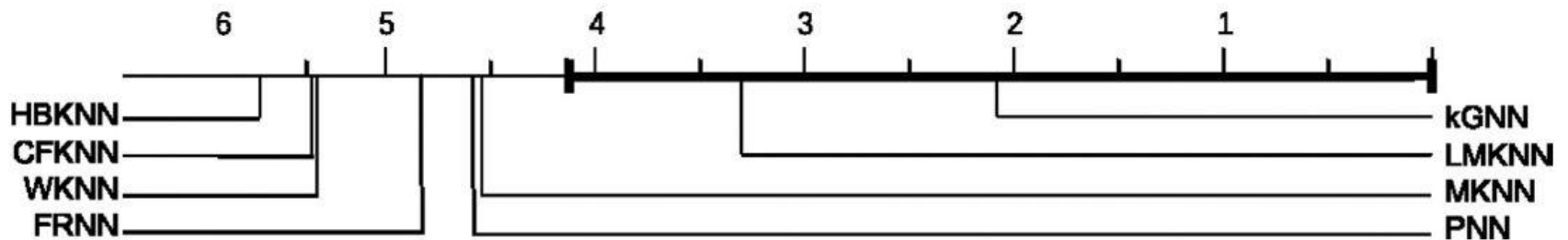
*Ranking stands the average ranking score of each classifier over twenty real-world datasets, where "1" is for the best and "8" is for the worst.

- 평균 순위 및 평균 오차율 모두 최상위
- 모든 데이터에 대해 최소 4순위 이상의 성능

5. 실험

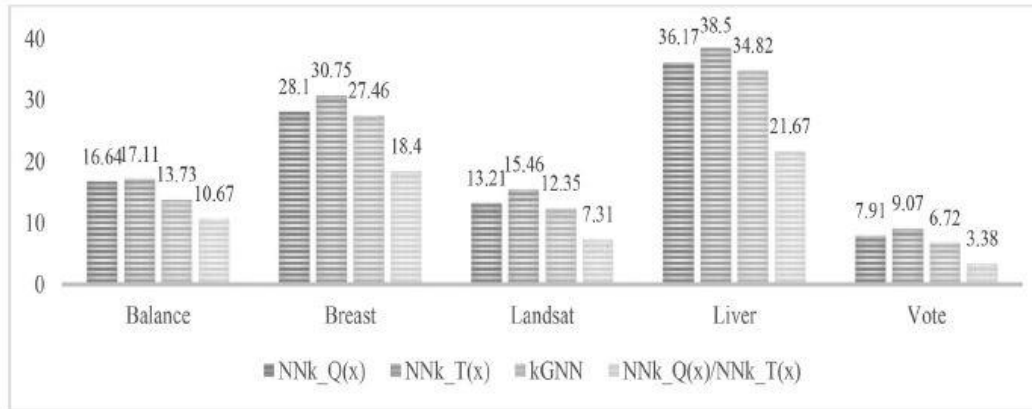
- k GNN과 다른 Classifier 비교
 - Bonferroni-Dunn test

$$CD = q_{\alpha} \sqrt{\frac{n_1(n_1 + 1)}{6n_2}} = 2.69 \sqrt{\frac{8 \cdot (8 + 1)}{6 \cdot 20}} = 2.08$$

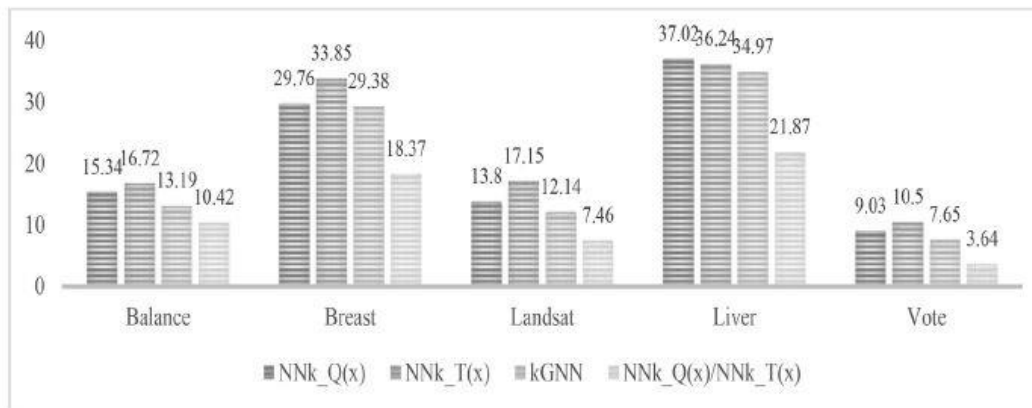


5. 실험

● k GNN의 특성



(a) Using L_1 -norm measure



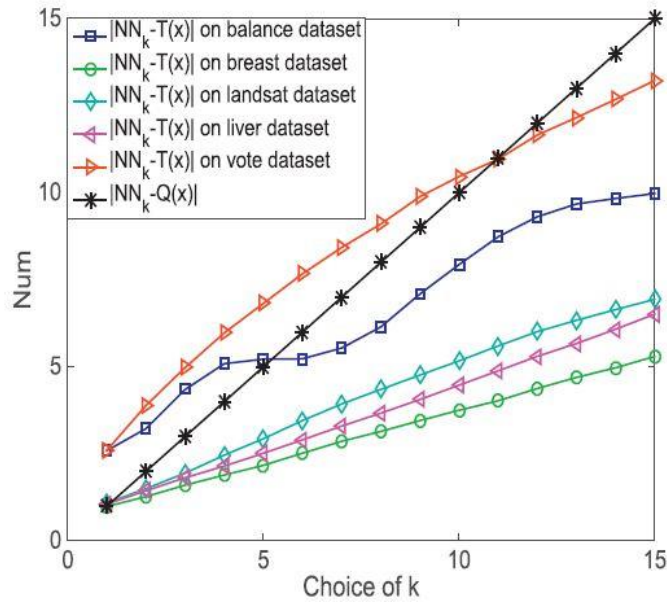
(b) Using L_2 -norm measure

Fig. 7. Complementary analysis of $NN_{k-Q}(x)$ and $NN_{k-T}(x)$ in k GNN in terms of the error rate (%) when $k = 5$.

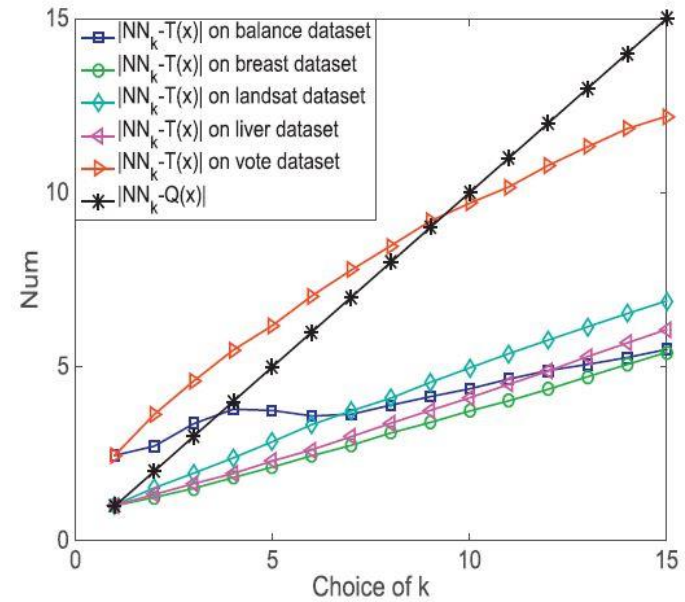
- $NN_{k-Q}(x)$ 보다 $NN_{k-T}(x)$ 가 대체적으로 오차율이 크다
- k GNN 방식은 각각을 따로 사용하는 경우보다 오차율이 작다
- k GNN 방식의 상호보완적 특성

5. 실험

● k GNN의 특성



(a) Using L_1 -norm measure



(b) Using L_2 -norm measure

Fig. 8. Comparison of $|NN_k-Q(x)|$ and $|NN_k-T(x)|$ when k varies from 1 to 15.

k 를 증가시킬수록 $|NN_k-T(x)|$ 의 비율이 줄어드는 경향

5. 실험

- k GNN과 k NN* 비교

Table 4

Comparison of k GNN and k NN* in terms of the error rate (%) using Euclidean norm (L_2 norm) and Manhattan norm (L_1 norm) when $k = 1, 5$, and 9 .

| Dataset | L_1 | | | | | | L_2 | | | | | |
|---------|---------|--------------|---------|--------------|---------|--------------|---------|--------------|---------|--------------|---------|--------------|
| | $k = 1$ | | $k = 5$ | | $k = 9$ | | $k = 1$ | | $k = 5$ | | $k = 9$ | |
| | k NN* | k GNN | k NN* | k GNN | k NN* | k GNN | k NN* | k GNN | k NN* | k GNN | k NN* | k GNN |
| Balance | 19.71 | 19.53 | 14.25 | 13.73 | 13.52 | 12.40 | 19.47 | 19.34 | 13.31 | 13.19 | 13.33 | 12.78 |
| Breast | 33.39 | 30.25 | 29.43 | 27.46 | 27.43 | 26.25 | 34.72 | 32.28 | 29.40 | 29.38 | 28.06 | 27.13 |
| Landsat | 12.95 | 12.92 | 12.46 | 12.35 | 13.15 | 13.00 | 13.69 | 13.48 | 13.46 | 12.14 | 13.85 | 13.78 |
| Liver | 40.43 | 40.39 | 36.04 | 34.82 | 35.63 | 34.91 | 40.33 | 40.26 | 36.24 | 34.97 | 36.20 | 34.95 |
| Vote | 8.25 | 8.18 | 8.03 | 6.72 | 8.97 | 7.17 | 9.43 | 9.43 | 8.84 | 7.65 | 9.55 | 8.27 |

동일한 k 에 대해 k GNN 방식이 더 낮은 오차율을 보인다

5. 결론

- training sample과 testing sample 각각의 neighborhood의 차이를 지적
- training sample과 testing sample 모두를 고려한 k GNN 방식 제안
- 다양한 실험을 통해 k GNN 방식의 성능, 타당성 및 특성을 보임