통계적 분류와 신경망을 이용한 필기 숫자 인식방법의 성능비교

오정환(2020517027): MLP, SDG 구현 및 사전데이터 입수 및 PCA 전처리 및 레포트작성
서지원(2020517004): KNN 및 의사결정트리 구현 및 이론 사전조사 및 레포트작성

요약

본 프로젝트에서는 필기 숫자 데이터들에 대한 인식률을 기준으로 통계적 분류 방법과 신경망 방법의 성능을 비교 분석하였다. 통계적 분류 방법으로 K-최근접 이웃기법(K Nearest Neighbor: KNN), 확률적 경사하강법(SGD), 의사결정트리 방법을 분석하였고, 신경망 방법으로는 다층퍼셉트론(MLP)을 선택하여 통계적 분류 방법과 비교하였다.

프로젝트는 Kaggle 데이터베이스를 사용하여 이루어졌으며, 주성분 분석(PCA) 방법으로 전 처리를 했을 때 학습속도 및 성능향상에 어떠한 영향을 미치는지에 대해서도 연구해 보았다. PCA 분석을 했을 때 인식률이 굉장히 떨어지는 것을 알 수 있었고, 실험결과 KNN이 가장 좋은 인식률을 보여주었다.

I. 서론

패턴 인식에서 활용되는 여러 기법들 중 대표적인 것이 통계적 분류 방법과 신경망 방법이다.

본 프로젝트에서는 필기 숫자에 대한 인식률을 기준으로 어떠한 기법이 더 우수한 성능을 보일지 비교 분석하였다. 통계적 분류 방법에서는 우선 입력 데이터가 지역적으로 얼마나 더 근사한가에 따라 분류해 학습하는 K-최근접 이웃기법(이하 KNN), 경사하강법에서 특정 데이터만을 뽑아 학습하는 확률적 경사하강법(이하 SGD), 직관적으로 분류할 수 있는 분석 방법인 의사결정트리를 채택하였다. 또한 신경망 기법으로는 가장 많이 응용되고 있는 기법 중 하나인 다층퍼셉트론(이하 MLP)을 선택하여 통계적 분류 방법들과 비교하였다. 마지막으로 주성분 분석(이하PCA)이 많은 특성 값을 가진 데이터의 학습 속도와 성능을 향상시키데 도움이 될 수 있는지 여부를 분석하였다.

II. PCA

분석하고자 하는 데이터가 많은 특성(feature)을 가지고 있다면 차원(dimension) 또한 증가하게 된다. 이러한 데이터를 활용해 학습을 하게 되면 복잡도가 올라가 결국 오버피팅의 위험이 커지게 된다. 따라서 차원축소 방법 중 하나인 PCA를 사용해 전 처리를 했을 때와 하지 않았을 때의 결과값을 비교해 우리가 연구하고자 하는 데이터를 학습하는데 더 좋은 방법을 찾고자 하였다.

[표1] PCA 처리 이후 학습 결과

PCA_components	Model	Data			
		5000	10000	20000	40000
2	KNN(K=2)	39.50%	39.60%	39.40%	38.73%
	KNN(K=3)	40.30%	39.30%	39.45%	39.70%
	SGD	6.80%	12.60%	16.70%	25.89%
	Decision Tree	28.70%	28.70%	29.03%	29.14%
	MLP	37.80%	36.30%	41.53%	41.11%
4	KNN(K=2)	56.50%	56.50%	58.35%	57.63%
	KNN(K=3)	59.40%	60.40%	60.68%	59.58%
	SGD	17.00%	13.20%	19.63%	21.95%
	Decision Tree	28.70%	28.80%	29.33%	29.21%
	MLP	58.30%	58.40%	56.78%	58.48%

표1은 데이터를 PCA를 이용해 전처리 후 각각의 분석 방법으로 학습한 결과이다. PCA로 전 처리를 하게 되면 어떠한 분석 방식으로도 굉장히 낮은 인식률을 보이는 것을 알 수 있다. PCA는 기존 데이터를 변환해 주성분으로 추출하는 방식이다. 하지만 필기 숫자 패턴 인식은 원래 데이터의 각 열의 값이 중요하기 때문에 기존 데이터를 변환해야 하는 PCA는 본 프로젝트의 데이터 학습 속도 및 성능 향상에 도움이 되지 않는다는 것을 알 수 있었다.

III. 통계적 분류 방법

통계적 분류는 데이터를 통계학 법에 의해 분류하는 기계학습의 과정이다.

3.1 KNN

KNN은 특정 공간 내에서 입력값과 제일 근접한 K개의 요소를 찾아, 더 많이 일치하는 것으로 분류하는 방법이다. 최적의 K값을 찾기 위해 K를 각각 2와 3일 때를 비교하여 학습하였다. 그 결과 K는 3일 때 더 좋은 분류 결과가 될 수 있음을 알 수 있었다.

3.2 SGD

경사하강법이란 함수의 기울기(경사)를 구하여 기울기가 낮은 쪽으로 계속 이동시켜서 극값에 이를 때까지 반복시키는 방법인데, SGD는 무작위로 추출한 하나의 샘플데이터에 대한 기울기를 계산해 적용하는 방법이다.

3.3 의사결정트리

의사결정트리(Decision Tree)는 각 데이터들이 가진 속성들로부터 패턴을 찾아내서 분류 학습을 시행한다.

IV. 신경망 분석 기법

신경망 분석 기법 중 대표적으로 자주 사용되는 MLP를 이용했다. MLP는 입력층과 출력층 사이에 하나 이상의 중간층(은닉층)이 존재하는 신경망이다.

V. 실험 결과 및 결론

Kaggle 데이터베이스 상의 필기 숫자에 대한 인식 실험을 통해 지금까지 설명한 통계적 분류 방법 및 신경망 기법의 성능을 비교하였다. 학습 훈련용으로 총 4,200자를 사용하였고, 785개의 열이 있다. “label”이라는 첫 번째 열은 사용자가 그린 숫자이고, 나머지 열은 이미지 픽셀 값이다. 사용된 각 이미지의 높이는 28 픽셀, 너비는 28 픽셀이며 총 784 픽셀이었다. 각 픽셀에는 해당 픽셀의 밝기 또는 어둡기를 나타내는 단일 픽셀 값이 있으며 숫자가 높을수록 어두워졌다.

데이터가 많은 것으로 판단되어 PCA분석을 통해 정제하려 했으나 PCA분석은 기존 컬럼의 값을 변화시키기 때문에 오히려 오리지널 데이터보다 정확도가 현저히 떨어졌다. KNN은 K값을 늘리면서 정확도를 향상시키

려 노력했으며 PCA 차원의 인자도 숫자를 늘려가며 정확도를 높여가려고 시도했다. 본 프로젝트 연구 결과 파악할 수 있었던 것은 크게 4가지다.

첫째, 데이터 개수가 각 모델에 주는 정확성은 얼마나 영향을 미치는가?

- 데이터 개수가 많아지면서 정확도는 전반적으로 정확도가 높아지는 현상을 발견할 수 있었다.

둘째, 수업시간에 배운 의사결정트리가 숫자 인식을 잘 할 수 있을지 연구했으나 컬럼수가 많은 숫자픽셀에 의사결정트리를 적용한다는 것은 정확도가 높지 않았다. 하지만 데이터 개수가 많아지면서 좀더 정교해 지는 것을 알 수 있었다.

셋째, SGD는 네트워크에서 내놓는 결과 값과 실제 값 사이의 차이를 정의하는 Loss Function의 값을 최소화하기 위해 기울기를 이용하는 것인데 PCA를 사용하면서 원본데이터를 손실하는 순간 정확도가 현저하게 낮아지는 것을 발견할 수 있었다.

넷째, 차원을 줄이면서 최적의 해를 찾는 PCA를 통해 좀 더 정확도가 높은 모델을 만들려 했지만 숫자 인식 데이터는 적합한 예제가 아니었다.

[표2] 분석 모델별 결과

Model	Data			
	5000	10000	20000	40000
KNN(K=2)	92.70%	92.70%	95.13%	96.21%
KNN(K=3)	91.70%	91.70%	96.13%	96.67%
SGD	86.90%	86.90%	86.73%	85.43%
Decision Tree classifier	33.40%	33.40%	33.73%	34.27%
MLP classifier Accuracy	88.10%	87.40%	93.55%	95.76%

결론적으로 필기 숫자 인식에는 KNN과 MLP가 가장 적합하다는 것을 도출할 수 있었다.

참고문헌

[1] *순환신경망을 이용한 한글 필기체 인식 (김병희, 장병탁), 2017*

[2] *SVM 분류기를 이용한 필기체 숫자인식 (박증조, 김경민), 2007*

[3] *Deep Big Multilayer Perceptrons For Digit Recognition (Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmid)*

[4] *패턴인식의 원리 (이성환, 홍릉과학출판사), 1994, 3~34P*