SVM 실제 사용 사례 조사

4조 (이희구, 유제우)

1. SVM을 이용한 출입 관리 시스템

그림1 Haar-like feature를 이용한 얼굴 탐지

- 얼굴 인식을 수행하기 위해서는 먼저 CCTV에서 촬영된 동영상의 프레임으로부터 얼굴 영역을 탐지해야 한다. 얼굴 탐지 방법으로 Haar-like feature 방법을 이용한다. Haar-like feature방법은 기존의 얼굴 탐지 방법들에 비해 얼굴 탐지 속도에서 뛰어난 성능을 보인다고 평가되고 있다. Haar-like feature 방법으로 탐지된 얼굴 영역은 사각형의 형태로 표현되며, 탐지된 얼굴 영역의 이미지는 얼굴 인식 시스템의 입력 데이터로 사용된다.

- 얼굴 인식의 벡터 표현으로 NMF(non-negative matrix factorization)를 사용한다. 입력 공간의 얼굴 영상들로부터 NMF에 의해 유도된 특징 공간의 기저 벡터는 눈, 코, 입 등과 유사한 부분 영상 형태를 보이며, 기저 벡터의 선형 결합으로 얼굴 영상을 나타낼 수 있다. NMF에 의해 표현된 특징 값은 영상 내 객체들의 공간적 정보를 포함하고 있으므로 이미지의 전체적인 정보만을 포함하고 있는 SVD나 PCA에 의한 특징 값보다 얼굴 인식에 적합하다고 알려져 있다.

- 얼굴의 빠르고 정확한 인식을 위하여 혼합 계층형 SVM (HHSVM: hybrid hierarchical SVM)이 사용된다. 계층적 다중 클래스 SVM과 단일클래스 SVM을 결합한 구조로써, 이진 트리를 응집 계층형 클러스터링 알고리즘을 이용하여 구성함으로써, 보다 빠른 시스템의 구성을 보장한다. 또한, 최상위 노드에 단일 클래스 SVM을 배치함으로써, 범죄자 후보의 여부를 빠른 속도로 판단한다. 결국, 범죄 자가 아닌 일반인에 대한 불필요한 연산을 줄임으로써 시스템의 부하를 경감하여 실시간 응용에 적합하도록 설계되었다.

- 분류해야 할 n개 클래스가 있을 경우, 기존의 다중 클래스 SVM에서는 최소 n개의 분류기를 통해 분류하였으나, 계층형 SVM 모델에서는 평균 log2n개의 분류기를 통해 클래스를 결정할 수 있다. 즉, 분류기가 이진트리 구조로 구성되면 매우 빠른 검색이 가능하게 된다. k-means(k=2) 알고리즘을 이용하여 이진트리를 구성하였으나, 이진트리를 구성할 때 k-means 알고리즘을 매 분기마다 수행해야만 하는 불편함이 있다. 반면, 응집 계층형 클러스터링 알고리즘을 사용하면 한 번의 클러스터링으로 이진 트리 구조를 생성할 수 있다. 이진 클래스 SVM을 이용한 분류기는 기본적으로 전체 공간을 2개의 공간으로 분할한 후 두 개의 공간 중에 더 가까운 공간 즉 클래스로 분류하므로, 학습되지 않은 새로운 클래스의 데이터에 대하여 새로운 데이터임을 알리지 못하고, 하나의 클래스로 강제 분류하게 된다. SVM은 선형 및 비선형 패턴을 구별하는 데 강력한 모델로 알려져 있으며, 얼굴 인식과 같은 복잡한 작업에 적합하고, 이상치에 상대적으로 덜 민감하므로, 얼굴 영상에서 노이즈나 이상치가 포함되어 있을 때에도 더 좋은 성능을 제공할 것으로 예상된다. 그렇기 때문에 SVM을 사용하지 않았을 때보다 높은 정확도를 보여준다.

2. SVM을 활용한 텍스트 분류

- SVM(Support Vector Machine)을 활용하여 텍스트 분류 작업을 수행하는 방법과 그 성능에 대해 다룬다. 텍스트 분류는 문서, 리뷰, 뉴스 기사, 소셜 미디어 게시물 등과 같은 텍스트 데이터를 특정 범주 또는 레이블에 할당하는 작업입니다.

▪ 스팸 필터링

– 이메일, 메시지, 댓글 등의 텍스트 데이터를 분류하여 스팸 여부를 판별하는데 SVM을 사용한다. SVM은 스팸과 비스팸(일반) 메시지를 분류하는데 효과적으로 활용됩니다.

▪ 감정 분석

– 소셜 미디어 댓글, 제품 리뷰, 뉴스 기사 등에서 텍스트 데이터를 분석하여 긍정적, 부정적, 중립적인 감정을 분류한다. 이를 통해 제품 평가, 브랜드 감성 분석, 고객 의견 모니터링 등에 활용된다.

▪ 뉴스 분류

– 뉴스 기사를 정치, 경제, 스포츠, 엔터테인먼트 등의 카테고리로 분류하는 작업에 SVM을 활용한다. 뉴스 집합을 자동으로 정렬하고 필요한 정보를 추출하는 것이 가능하다.

▪ **법률 문서 분류**

- 변호사 사무실에서는 SVM을 사용하여 다양한 법률 문서를 분류한다. 예를 들어, 판례, 계약서, 소송 서류 등을 자동으로 정리하고 관리할 수 있다.

▪ **의료 진단**

- 의료 분야에서는 환자 기록, 의료 보고서 등의 텍스트 데이터를 SVM을 사용하여 질병, 증상, 검사 결과 등으로 분류한다. 이를 통해 의사들은 빠르게 정보를 검색하고 진단을 지원받을 수 있다.

▪ **카테고리 기반 광고 배치**

- 온라인 광고 플랫폼에서는 텍스트 내용을 분석하여 해당 광고 카테고리에 맞게 광고를 배치하는 데 SVM을 사용한다. 사용자 콘텐츠와 일치하는 광고를 표시하여 광고 효과를 극대화한다.

▪ **자연어 질의**

- 검색 엔진 및 가상 비서(AI 어시스턴트)에서는 사용자의 자연어 질의를 이해하고 적절한 답변을 생성하기 위해 SVM과 같은 텍스트 분류 모델을 활용한다.

▪ **금융 서비스**

- 금융 분야에서는 기업 보고서, 금융 뉴스, 주식 시장 댓글 등에서 정보를 추출하고 금융 이벤트 및 트렌드를 예측하기 위해 SVM을 활용한다.

- 텍스트 데이터를 수집하고, 필요한 경우 전처리를 수행한다. 이 단계에서는 텍스트 데이터를 토큰화하고, 불용어(stopwords)를 제거하며, 특수 문자나 숫자를 정제하는 등의 단계가 포함된다. 텍스트 데이터를 기계 학습 모델에 입력할 수 있는 형태로 변환한다. 주로 TF-IDF와 같은 특징 추출 방법을 사용한다. 이렇게 추출된 특징 벡터는 각 문서의 단어 빈도와 중요성을 나타낸다. 데이터를 훈련 세트와 테스트 세트로 분할한다. 훈련 세트로 모델을 훈련하고, 테스트 세트로 모델의 성능을 평가한다. SVM 모델을 선택하고, 필요한 경우 하이퍼파라미터를 조정한다. SVM에서는 커널 종류와 C 매개변수 등을 조정할 수 있다. SVM 모델을 훈련 데이터에 맞춘다. 모델은 훈련 데이터의 특징 벡터를 입력으로 받고, 클래스(카테고리)를 예측한다. 테스트 세트를 사용하여 모델의 성능을 평가한다. 일반적인 평가 메트릭으로는 정확도, 정밀도, 재현율, F1 점수 등이 사용된다.

- SVM을 사용함으로써 텍스트 분류 작업의 정확도가 향상된다. SVM은 데이터를 비선형 경계로 분류하거나 고차원 공간으로 매핑하여 복잡한 패턴을 잘 인식할 수 있다. 따라서 SVM을 사용하면 레이블 또는 범주 간의 구별이 더 명확하게 이루어진다.

- SVM은 일반적으로 텍스트 데이터에서 발생하는 다양한 패턴을 잘 일반화할 수 있다. 이는 다양

한 문서 유형, 언어, 스타일 등을 다룰 때 유용하다. 일반화 능력의 향상으로 인해 모델은 새로운 데이터나 문서에 대해서도 더 정확한 예측을 수행할 수 있다.

3. SVM을 활용한 생물정보학

▪ **유전자 발현 분류**

- SVM은 유전자 발현 데이터를 분석하여 다양한 종류의 종양, 질병 또는 조건을분류하는 데 사용됩니다. 이를 통해 특정 유전자 패턴이 특정 질병 또는 조건과 연관되어 있는지를 확인하고, 진단 및 치료에 도움이 되는 유전자 마커를 식별하는 데 사용됩니다.

▪ **단백질 분류**

- SVM은 단백질 서열 분류에도 활용됩니다. 단백질 서열을 분류하여 기능, 구조 또는상호 작용을 예측하고, 단백질 기능을 이해하고 해석하는 데 도움을 줍니다.

▪ **질병 예측 및 예방**

- SVM은 유전자, 단백질, 화학 구조 등의 데이터를 기반으로 질병 예측 및 예방에 활용됩니다. 예를 들어, 암 진단에서는 SVM을 사용하여 조직 샘플에서 암의 발생 여부를 예측하고, 개별 환자에게 맞춤형 치료 방법을 제안하는 데 사용됩니다.

▪ **바이오인포매틱스**

- SVM은 바이오인포매틱스 분야에서 유전체 및 단백질 서열 분석, 유전체 어셈블리, 비슷한 서열 찾기, 진화 및 분류 작업에 널리 사용됩니다. 서열 데이터를 분류하고 예측하여 생물학적 의미를 추론하고 식별하는 데 유용합니다.

- 연구자는 데이터를 수집하고 필요한 형식으로 전처리를 한다. 예를 들어, 유전자 발현 데이터의 경우 각 유전자의 발현 수준이 측정된 특성 행렬로 표현된다. 데이터는 학습 데이터와 테스트 데이터로 분할된다. 학습 데이터는 SVM 모델을 학습시키는 데 사용되고, 테스트 데이터는 모델의 성능을 평가하는 데 사용된다. 각 데이터 포인트는 특성 벡터로 표현된다. 이 벡터는 데이터의 특징이나 속성을 나타낸다. SVM은 데이터를 고차원 특성 공간으로 매핑하여 비선형 관계를 다룰 수도 있다. SVM은 데이터를 분류하기 위한 최적의 결정 경계를 찾는 것이 목표이다. 학습 데이터를 기반으로 SVM 모델을 학습시키는데, 이때 각 데이터 포인트를 고차원 공간에서 서로 구분하는 결정 경계를 찾는다. 이 결정 경계는 Support Vectors(지지 벡터)라고 하는 일부 학습 데이터 포인트 주변으로 형성된다. SVM 모델에는 하이퍼파라미터가 있으며, 이러한 매개 변수

를 조정하여 모델의 성능을 최적화한다. 대표적으로 커널(kernel) 함수의 종류나 규제 파라미터(C값)를 조절하는데 주로 사용된다. 학습된 SVM 모델을 사용하여 테스트 데이터를 분류하고 예측한다. 예측 결과는 모델의 성능을 평가하는 데 사용된다. 일반적인 평가 지표로는 정확도, 정밀도, 재현율, F1 점수 등이 있다. SVM 모델을 통해 얻은 결과는 연구자나 의사들에게 중요한 정보를 제공하며, 예를 들어 유전자 발현 데이터에서 어떤 유전자가 특정 질병과 관련되어 있는지 또는 단백질 분류에서 어떤 서열이 특정 기능을 수행하는지를 이해하는 데 도움이 된다.

- SVM을 사용했을 때가 사용하지 않았을 때보다 비선형 데이터 관계도 잘 인식할 수 있어서 다양한 데이터 패턴을 포착할 수 있고, 데이터의 다양한 특성과 상호 작용을 고려하여 더 정확한 예측을 할 수 있다. 모델의 정확도와 일반화 능력이 향상되어, 더 신뢰할 수 있는 결과를 얻을 수 있다. 이상치나 노이즈에 대한 강건성이 높아져 데이터의 품질이 낮거나 이상치가 있는 경우에도 더 정확한 예측을 할 수 있다.

4. SVM을 활용한 이미지 분류

- SVM을 활용한 이미지 분류는 이미지 데이터를 특징 벡터로 변환하고, SVM 모델을 학습하여 이미지를 다른 클래스로 분류하는 기술입니다. 이를 통해 객체 인식, 의료 진단, 보안 검색과 같은 다양한 응용 분야에서 이미지를 정확하게 분류할 수 있습니다.

▪ **고차원 특징 공간 다룸**

- SVM은 이미지를 고차원 특징 공간으로 매핑하여 복잡한 패턴을 잡아냄으로써 비선형 문제를 해결할 수 있다.

▪ **서포트 벡터 활용**

- SVM은 결정 경계 주변의 일부 데이터 포인트만을 사용하는 서포트 벡터를 활용하여 모델을 구축하므로, 메모리 사용 효율성이 뛰어나고 과적합을 방지할 수 있다.

▪ **커널 기법 적용**

- 커널 기법을 통해 비선형 관계를 효과적으로 모델링하고 다양한 이미지 특징을 다룰 수 있다.

▪ **일반화 능력 향상**

- SVM은 일반화 능력이 뛰어나기 때문에, 학습 데이터에 없는 새로운 이미지에 대해서도 높은 정확도를 유지한다.

- SVM을 사용하지 않았을 때, 이미지 분류는 주로 단순한 통계 기반 방법으로 수행되며, 이로 인

해 다음과 같은 문제들이 발생한다. 첫째, 이미지 내의 복잡한 패턴이나 다양한 특징을 인식하기 어려워 정확한 분류가 어렵다. 둘째, 과적합 문제가 발생할 가능성이 높아, 학습 데이터에만 적합한 모델을 생성하고, 새로운 이미지에 대한 일반화 능력이 부족하다. 셋째, 복잡한 이미지 데이터를 분류하거나 객체 인식을 수행하는 데 한계가 있다. 마지막으로, 정확도와 신뢰성이 낮아, 의료 진단, 보안 검색 및 자율 주행 자동차와 같이 정확한 이미지 분류가 필요한 응용 분야에서 한계가 있다.

- SVM을 사용한 이미지 분류는 위 문제들을 극복하고 더 나은 성능을 제공한다. SVM은 이미지 분류 작업에서 복잡한 패턴 및 다양한 특징을 효과적으로 다룬다. 또한 서포트 벡터와 커널 기법을 사용하여 일반화 능력을 향상시키며, 새로운 데이터에 대한 예측 정확도가 높아진다. 과적합을 방지하면서도 높은 정확도와 신뢰성을 제공하므로, 객체 인식, 의료 진단, 보안 검색과 같은 응용 분야에서 뛰어난 성능을 보인다.
- 결론적으로 SVM을 사용한 이미지 분류는 이미지 분류 작업에서 정확도와 일반화 능력을 향상시키며, 복잡한 이미지 패턴을 인식하는 데 효과적이다. SVM을 사용하지 않았을 때는 정확도와 성능이 한계에 도달할 수 있다.