트레이닝 데이터 감소를 위한 병렬 평면 기반의 Support Vector Machine

이태호O, 김민우*, 이병준*, 김경태**, 윤희용*
성균관대학교, 정보통신대학 전자전기컴퓨터공학과O
성균관대학교, 정보통신대학 전자전기컴퓨터공학과*
성균관대학교, 소프트웨어대학 소프트웨어학과**
e-mail: {leetaeho, kimmw95, byungjun}@skku.eduO*, kyungtaekim76@gmail.com**, youn7147@skku.edu*

Support Vector Machine Using Parallel Hyperplane for Reduction of Training Data

Tae-Ho LeeO, Min-Woo Kim*, Byung-Jun Lee*, Kyung-Tae Kim**, Hee-Yong Youn*
Dept. of Electrical and Computer Engineering, Sungkyunkwan UniversityO
Dept. of Electrical and Computer Engineering, Sungkyunkwan University*
Dept. of Software, Sungkyunkwan University**

● 요 약 ●

SVM (Support Vector Machine)은 견고성으로 인해 다양한 분류 문제에 적용 할 수 있는 효율적인 기계 학습 기술이다. 그러나 훈련 데이터의 수가 증가함에 따라 시간 복잡도가 급격히 증가하므로 대규모 데이터 세트의 경우 SVM이 비실용적이다. 본 논문에서는 SVM을 사용하여 중복 된 학습 데이터를 효율적으로 제거하는 새로운 병렬 평면(Parallel Hyperplane) 기법을 소개한다. 제안 기법에서 PH는 재귀 적으로 형성되는 반면 PH의 외부에 있는 데이터 포인트의 클러스터는 매 반복마다 제거된다. 시뮬레이션 결과 제안 기법은 기존의 클러스터링 기반 감축 기법과 SMO 기법에 비해 학습 시간을 크게 단축시키면서 데이터 축소 없이 분류의 정확성을 높일 수 있음을 확인 하였다.

키워드: SVM (Support Vector Machine), 병렬 평편(Parallel Hyperplane), 데이터 감소(Data Reduction)

I. Introduction

SVM(Support Vector Machine)은 선형 및 비선형 분리 데이터 모두에 유용한 실용적인 접근법이다. 비선형으로 분리 가능한 데이터에 사용되는 주요 개념은 데이터 포인트를 선형으로 분리 할 수 있는 커널 함수를 사용하여 저 차원 데이터 포인트를 고차원 공간으로 매핑하는 것이다. SVM의 주요 작업은 데이터 간의 분리 공간을 최대화하는 최적의 초평면을 찾는 것이다. 하지만 다양한 분류 문제로 널리 받아 들여졌지만 교육 데이터를 처리하기 위해서는 많은 양의 계산과 메모리가 필요하다. 이때 중복 데이터 포인트라고하는 초평면의 구성에 영향을 미치지 않거나, 거의 영향을 미치지 않는 훈련 데이터의 수를 줄임으로써 문제를 완화 할 수 있다. Supprot Vector(SV)라고 불리는 소그룹의 훈련 샘플만이 SVM에서 초평면의 생성에 영향을 미친다. 따라서, SV와 관련이 없는 훈련 표본은 결정 기능에 영향을 미치지 않고 제거할 수 있다. 여기서 핵심은 주어진 교육 데이터 세트 중에서 중복 데이터 포인트를 정확하고 효율적으로 식별하는 것이다. 현재 SVM을 이용한 훈련 계산 오버헤드를 줄이기 위한 다양한 접근법이 제안되었다[1]. 그 중 클러스터링 알고리즘을 SVM과 결합하는 것은 SVM의 복잡성을 줄이기 위해 사용되는 일반적인 방법 중 하나이다. 본 논문에서는 분류 정확도를 떨어트리지 않으면서 트레이닝 데이터의 양을 크게 줄이기 위해 Parallel Hyperplane(PH)이라는 새로운 개념을 제안한다.

II. Preliminaries

1. Related works

계층적 클러스터링, 퍼지 클러스터링 등과 같은 SVM의 훈련 데이터를 줄이기 위한 다양한 접근법이 존재하지만, 클러스터와 초평면의 조작에 대해서는 거의 관심을 기울이지 않았다.

III. The Proposed Scheme

근사 초평면이라고 부르는 것은 클러스터의 중심에 기초하여 만들어지며, PH는 평행하고 클러스터 중심의 중심을 통과한다. 데이터 포인트가 PH와 근사 초평면 사이에 위치하지 않는 클러스터는 SV에

포함될 수 없기 때문에 학습 데이터 세트에서 제거한다. 그런 다음 나머지 클러스터로 새 PH를 구성하고 축소가 불가능할 때까지 프로세스가 반복한다.

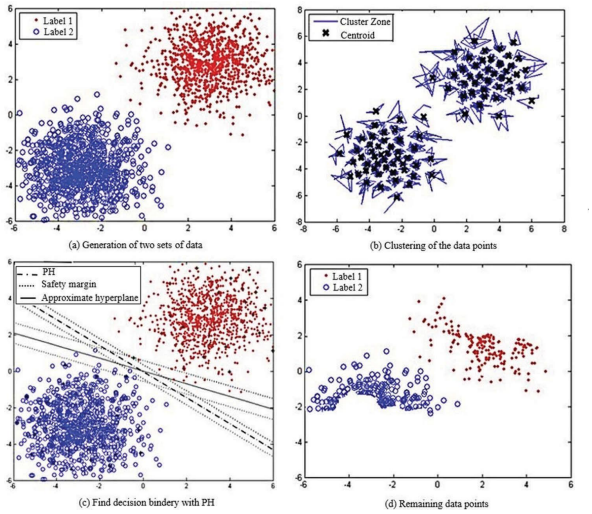

Fig. 1. The steps of the proposed PH scheme with normally distributed data

제안 기법의 유효성을 검증하기 위해 클러스터 기반 알고리즘 및 SMO 알고리즘과 비교 하였을 때 교육 시간과 정확성, 분류의 정확도를 떨어트리지 않으면서 교육 시간을 크게 감소시킨다는 것을 알 수 있다. 이는 제안 기법이 중복 데이터 포인트를 생략함으로써 SV를 보존하므로 초평면의 구성에 해를 끼치지 않기 때문이다. 또한 표준 편차가 비교적 큰 데이터 집합과 특정 지점까지 많은 수의 클러스터에 대해 더 효과적인 것으로 나타났다.

IV. Conclusions

본 논문에서는 SVM의 학습 시간을 줄이기 위해 학습 데이터 세트에서 중복 데이터 포인트를 효과적으로 제거하는 PH 기법을 제안했다. 제안 기법에서 k-mean 클러스터링 알고리즘을 사용하여 주어진 데이터 포인트를 다른 클러스터로 나눈 다음 잠재적으로 벡터를 지원하지 않는 클러스터의 데이터 포인트를 학습 데이터 세트에서 제거한다. 제안 기법은 상당한 양의 중복 데이터 포인트를 제거하고 분류의 정확도에 영향을 미치지 않으면서 결국 교육 시간을 단축시킨다는 것을 확인했다. 또한 제안 기법이 기존의 클러스터 기반 기법보다 훨씬 빠르다는 것을 알 수 있다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송연구 개발 사업(No. 2016-0-00133, 초연결 IoT 노드의 군집 지능화를 통한 Edge Computing 핵심 기술 연구), SW중심대학지원사업(2015-0-00914), 한국연구재단 기초연구사업(No.2016R1A6A3A11931385, 실시간 공공안전 서비스를 위한 소프트웨어 정의 무선 센서 네트워크 핵심기술 연구, 2017R1A2B2009095, 실시간 스트림 데이터 처리 및 Multi-connectivity를 지원하는 SDN 기반 WSN 핵심 기술 연구, 2019R1I1A1A01058780, 머신러닝 기술을 사용한 SDN기반 무선센서네트워크의 효율적 관리), BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] N. Cristianini and S. T. John. “An introduction to support vector machines and other kernel-based learning methods”, Cambridge university press, 2000.

[2] J. Cervantes, X, Li and W. Yu, "Support vector machine classification based on fuzzy clustering for large datasets", Mexican International Conference on Artificial Intelligence, pp.572-582, 2006.