Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

유전적 특징선택에 관한 연구

A Study on Genetic Feature Selection

한명묵
Myung-Mook Han

경원대학교 IT대학 컴퓨터소프트웨어학과
E-mail: mmhan@kyungwon.ac.kr

요 약

많은 분야에서 최적의 기준을 바탕으로 특징들의 부분집합을 선택하는 문제들이 핵심 요소로 작용하고 있다. 다양한 특징들의 부분집합 중에서 가능한 한 가장 성능이 우수한 특징들의 부분집합을 선택하기 위해서는 특징선택 방법이 알고리즘과 적용분야들을 고려해야한다. 이 논문에서는 특징선택을 위해서 서로 다른 두 종류의 최적화 문제를 탐색하는 방법을 제안하고, 그 결과를 실험으로 보여준다.

키워드 : 유전자알고리즘, 특징 선택. 하이브리드 시스템, 최적화 방법, classification.

1. 서 론

과거 40여년동안 특징선택을 위한 방법들을 개선하기 위해서 상당한 연구가 진행되어 왔다. 이 설계의 질은 다양한 특징들의 유용도, 성능, 그리고 복잡도에 관계가 있다. 특징선택방법들은 무엇으로 평가하는지에 따라 필터(filter) 접근과 래퍼(wrapper) 접근으로 나눌 수 있다 [1]. 필터 접근은 분류과정 전에 인스턴스들 사이의 어떤 측정된 거리를 바탕으로 특징들의 부분집합을 선택한다. 래퍼 접근은 분류 과정 중에서 분류의 결과를 바탕으로 특징들의 부분집합들을 선택한다.

유전자 알고리즘(Genetic Algorithm:GA)은 자연의 적자생존과 진화의 방법을 모방한다. 특별히 최적화 문제를 풀기 위해서 보답(payoff)을 올리거나 값을 내리는 확률적 탐색 기법을 이용한다. 또한 GA는 다중모드(multimodal) 탐색 도메인에서 전역 해를 최적으로 발견하는데 높은 확률을 가지고 있다. 특징 부분집합 선택은 모든 가능한 부분집합 중에 분류 성능을 최대화하는 특징들의 부분집합을 선택하는 과정으로 정의된다. 매우 큰 문제에서는 탐색해야 할 탐색 공간은 매우 크며, 특히 특징 선택 문제는 다중 기준(multicriteria)과 제약(constraint)을 갖는 최적화 문제를 나타낸다. 이러한 점은 GA가 특징 선택 문제에 적용될 수 있는 것을 보여준다[2].

본 논문에서는 필요한 특징들만을 GA를 활용해서 선택하는 방법을 제안한다. 이 방법은 세 단계를 거치면서 진행이 되며 무엇을 최적화 하느냐에 따라 적응함수를 다르게 정하지 않고 적용할 수가 있어서 다양한 도메인에서 최적의 특징들의 부분집합을 선택해서 사용할 수 있는 장점이 있다.

논문은 다음과 같이 구성된다. 2장에서 특징선텍에 관한 연구에 대해서 설명한다. 3장에서 제안하는 모델인 GA를 활용한 특징 선택 방법을 소개하고, 실험 결과와 결론을 4장에서 정리한다.

2. 특징 선택 문제

분류과정에서의 첫 번째 단계는 크고 원래의 특징 집합에서 작은 특징의 부분 집합을 선택하는 특징 선택이다. 인스턴스는 분류 알고리즘에 값 {v1, v2, ..., vn}을 {f1, f2, ..., fn} 특징 집합에 클래스 레벨 c와 함께 배정함으로써 기술된다. 만약에 함수 F가 학습되어져야 되면, c = F(v1, v2, ...vn)이다. 특징 선택은 분류 성능을 최대한으로 하면서 주어진 원래의 n 특징들에서 m의 유용한 특징들을 선택하는 것이다.

특징 선택 방법들은 평가하기 위해서 무엇을 사용하느냐에 따라 필터와 래퍼 접근 방법으로 나눌 수가 있다. 필터 접근 방법은 분류 과정 전에 인스턴스들 사이에 거리의 측정을 바탕으로 특징들의 부분 집합을 선택하는 것이다. 래퍼 접근 방법은 분류 과정 중에 분류의 결과를 바탕으로 특징의 부분집합을 선택한다.

통계, 지리학, 기계학습 등을 포함한 다양한 방법들을 통해 연구자들이 시도했던 특징 선택 방법들은 상당히 많이 존재한다.

통계학 방법에서는 전방과 후방 stepwise multiple regression(SMR)이 특징들을 선택하기 위해 사용된다. 전방 방법은 후방 방법에 비해서 복잡도가 덜하기 때문에 전방 방법이 주로 사용된다.

지리학 방법에서는 탐색 공간에서 인스턴스의 위치들이 결정트리를 위한 특징들을 선택하기 위해서 IDG 알고리즘에 보내진다. 다른 클래스로부터 경계 instance를 분리하는 규칙들은 보상을 받고, 같은 클래스로부터 경계 instance를 분리하는 규칙들은 벌칙을 받는다.

기계학습 방법에서는 sequential forward search(SFS), sequential backward search(SBS) 그리고 변형 방법들이 사용되어졌다. SFS는 빈 집합으로부터 시작해서 지역적으로 가장 나쁜 특징을 제거한다. 교사 훈련 알고리즘을 가지고 훈련시키는 뉴럴 네트워크 방법들은 오용 침입 탐지에 적용이 되고, 비교사 훈련 알고리즘을 갖고 훈련시키는 뉴럴 네트워크는 비정상 침입

탐지에 적용이 된다[3]. 퍼지 집합 이론을 활용해서 FuzzyARTMAP이 적용되고[4], 러프 집합 이론을 가지고 PRESET은 이진 특징들을 선택하기 위해서 특징 집합들의 종속성을 결정한다. GA는 특징집합을 집합에서 특징들의 존재 유무를 표시하는 1과 0의 비트스트링으로 코딩해서 특징 선택을 한다.

3. 유전적 특징 선택

3.1 관련 연구

특징 부분집합 선택은 모든 가능한 부분집합에서 분류 성능이 최대화하는 특징들의 부분집합들을 선택하는 과정이다. 넓은 범위의 문제에서 탐색해야 할 탐색 공간은 매우 클 수가 있다. 또한, 특징 선택은 여러 조건과 제약의 최적화 문제이다. 이러한 점이 GA가 특징 선택 문제에 적용하기에 적당하다.

Sklansky등은 전형적인 방법과 비교해서 GA의 우수성에 대한 결과를 제공하였다. 그들은 또한 GA가 NP-hard 문제들을 푸는데 중요한 대안임을 보여주었다. GA 기반의 방법은 Greedy 같은 탐색 방법과 비교되었다.

GA를 활용한 여러 방법들이 제안되었다. 그 중 대부분은 분류 알고리즘의 에러를 최소화해야 할 적응 함수로 사용했으며, 래퍼 접근 방법을 사용하였다. AQ15, ID3/C4.5, 그리고 K-nearest neighbor 분류 알고리즘 같은 여러 분류기들이 에러 비율을 평가하기 위해서 사용되어졌다. 뉴럴 네트워크와의 결합이 또한 패턴 분류와 IDS를 위한 시스템 구조를 설계하는데 제안되었다.

특징들의 부분집합을 탐색하기 위해서 각 개체는 통상 n 비트 스트링으로 코드될 수 있는데, 여기서 0 값은 특징 집합에서 제외 된 특징을 1 값은 특징 집합에서 포함되어 있는 것을 나타낸다.

3.2 하이브리드 유전적 특징 선택

분류 정확도를 향상시키기 위한 목적을 가진 많은 수의 알고리즘들이 특징 선택을 위해서 제안되어 왔다. 특징의 데이터 집합들은 통상 정보처리를 하는데 있어 불필요한 많은 특징들을 가지고 있다. 그러한 특징들은 분류과정에서 사용되지 않기 때문에, 상대적으로 중요한 특징들만 선택하는 것이 분류 정확도를 향상시키는데 기여한다.

다양한 특징 선택 알고리즘 중에, GA 방법은 성공적으로 개발되어왔다. 그러나, GA를 사용한 접근방법들은 무엇을 최적화하는지에 대한 고려없이 가장 작은 에러율을 가진 특징의 부분집합을 찾는다. M.Kudo와 J.Sklansky는 특징선택 알고리즘에 대한 방향을 제시했다. 그들은 최적화 목적을 바탕으로 세 그룹으로 나누었다. 한 그룹은 에러율이 최소화 되도록 주어진 특징 집합의 부분집합을 구한다. 두 번째 그룹은 에러율이 어떤 기준치보다 높지 않은 가장 작은 수의 특징들을 갖는 부분집합을 구한다. 마지막 그룹은 작은수의 특징부분집합하고 에러율사이의 조정된 것을 발견한다. 또한 알고리즘들은 다른 목적에 따라 다른 적응함수를 사용했다.
그러나 GA를 사용하는데 있어서 상황에 따라 서로 다른 적응함수를 사용하는 것은 불편할 뿐만 아니라 전체 문제에서 부분최적화로 될 수가 있다. 그러므로 이 논문에서는 특징의 부분집합중에서 최적의 부분집합을 찾는 하이브리드 방법을 제안한다.

제안하는 방법은 그림 1과 같이 세 단계를 거친다.

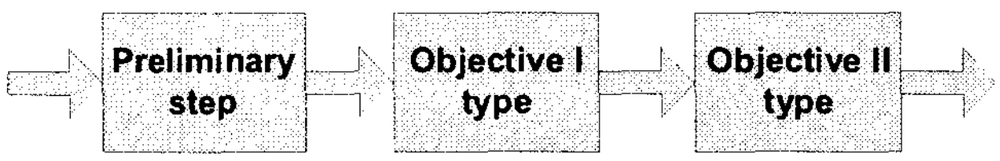

그림 1. 제안하는 방법의 구조
Fig. 1. The structure of proposed method.

4. 실험결과 및 결론

4.1 Preliminary step

특징선택 알고리즘은 criterion curve라는 곡선으로 표현될 수 있다. 곡선은 X축이 특징들의 수, Y축은 에러율 상에서의 점들의 집합을 연결한다. 특징선택문제는 monotonic 혹은 근사 monotonic이기 때문에 이 곡선에서 그림 2와 같이 최소 평가 값인 $Error_{min}$에서부터 a% 낮은 값을 그림 2와 같이 선정했다.

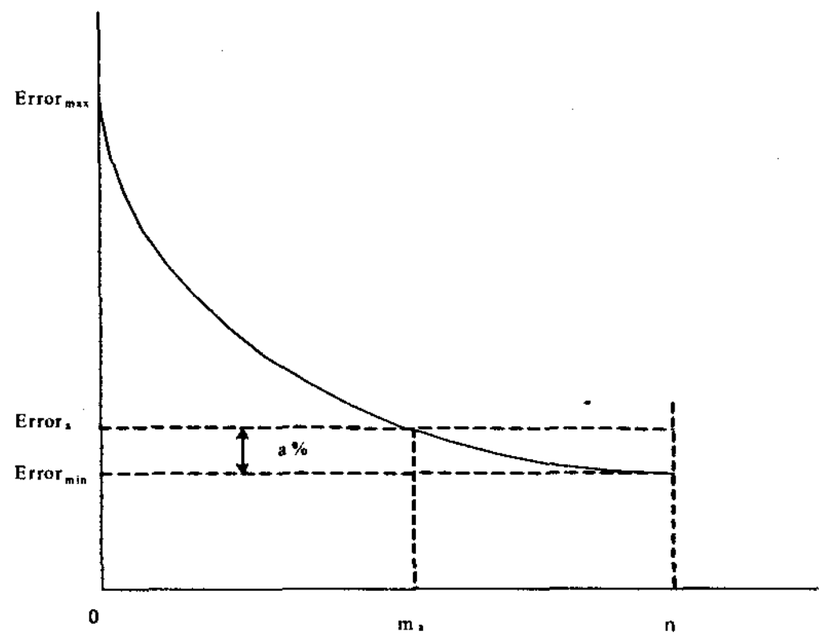

그림 2. 우선단계
Fig. 2. Preliminary step.

4.2 결론

본 논문에서 하이브리드 유전적 알고리즘 방법을 제안했다. 이 방법은 무엇을 최적화하는지에 따라 적응함수가 달라지는 문제를 제거했으며, 두 종류의 적응함수를 연결해서 사용했기 때문에 어떤 부분의 특징 선택문제에도 적용할 수가 있다. 현재 다른 방법들하고 비교연구를 위한 실험과 특정분야에서 적용할 수 있는 방법들에 관해서 연구하고 있다.

참 고 문 헌

[1] Langley, P., Selection of Relevant Features in Machine Learning, In Proc. of the AAAI Fall Symposium on Relevance, New Orleans, LA: AAAI Press, 1994.

[2] J.H.Holland, Adaptation in NAtural and Artificial Systems, Univ. of Michigan Press, Ann Arbor, Mich., 1975.

[3] A.J.Hoglund, K.Hatonen, and A.S.Sorvari, A computer host-based user anomaly detection system using the self-organizing map, om Proc. of the IEEE-INNS-ENNS Int.Joint Conf. on Neural Networks(IJCNN2000), Vol.5., pp.411-416, 2000.

[4] J.Cannady and R.C.Garcia, The application of fuzzy ARTMAP in the detection of computer network attacks, in ANN ICANN2001.