

정보이론에 기반한 Supervised, Unsupervised 피처 선택 방법론

이상근^o 장병탁
서울대학교 컴퓨터공학부
sklee^o@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Information-based Supervised and Unsupervised Feature Selection Methods

Sang-Kyun Lee^o and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요약

많은 변수(variable)와 피처(feature)를 포함하는 대규모 데이터에 기계학습 방법론을 적용하는데 있어 그 예측 성능을 향상시키기 위한 방법으로 피처 선택(feature selection) 기법이 활발히 연구되고 있다. 그러나 다른 연구를 위한 사전 데이터 분석 작업에 유용하게 사용될 수 있는 단순한 순위기반 피처 선택 방법론은 피처의 중요한 특성을 간과하는 경우가 많으며, 따라서 예측 성능의 향상을 기대하기 어렵다. 본 연구에서는 정보이론에 기반한 supervised 피처 선택 방법과 이것을 보완할 수 있는 unsupervised 피처 선택 방법을 제시했다. 서로 다른 특성을 가진 다섯 개의 데이터셋에 대해 실험한 결과, 제시된 방법이 기존 방법보다 나은 예측 성능을 보임을 확인했다. 또한 두 방법에서 얻어진 피처들을 결합해 사용할 경우 한가지 방법만으로 추출된 피처를 사용할 경우보다 나은 기계학습 성능을 보임을 확인했다.

1. 서론

최근 기계학습(machine learning) 연구에서 사용되는 데이터는 일반적으로 많은 변수와 피처(feature)를 포함하는데, 그 대표적인 예가 마이크로어레이(microarray) 기법을 통해 얻은 유전자발현데이터(gene expression data)와 문서분류(text categorization)에 사용되는 문서 데이터이다[1]. 마이크로어레이 데이터는 주로 환자들의 유전자 프로파일(gene profile)을 분석하여 환자와 정상인을 분류하는데 응용되는데, 수천 단위의 유전자를 포함하는 것이 일반적이다. 또한 문서분류 데이터에서는 하나의 문서가 단어의 벡터로 표현되는데, 보통 수십만 단위의 단어를 포함한다.

이렇게 많은 피처를 포함하는 데이터에서 특정 피처 집합을 선택하여 얻을 수 있는 효과는 다양하다. 즉 데이터 시각화(visualization)을 용이하게 하고, 데이터의 이해를 증진시키며, 계산 시간이나 저장공간 필요량을 줄여주며, 검색 차원(dimension)을 감소시켜 기계학습의 예측 성능을 높이는 것이다. 이 중에서도 본 연구는 기계학습의 예측 성능 향상에 목적을 두고 있다.

예측 성능 향상을 위한 피처 선택(feature selection)에서는 일반적으로 많은 계산 시간을 필요로 하는 복잡한 방법론이 좋은 성능을 보인다[2]. 가장 좋은 피처들을 추출하기 위해서는 물론 이러한 방법론을 사용해야겠지만, 좀더 빠르게 피처를 추출하는 방법 역시 다른 연구를 위한 사전 데이터 분석 차원에서 중요한 의미를 갖는다. 하지만 간단하면서도 빠른 방법들, 특히 상호정보량의 순위에 따라 피처를 선택하는 방법은 피처의 다른 중

요한 측면들을 간과하는 경우가 많으며, 따라서 이러한 측면을 보충할 수 있는 방법론이 필요하다.

본 연구에서는 정보이론에 기반하여 각 피처의 중요도를 평가하는 supervised, unsupervised 피처 선택 방법을 제시했다. 또한 두 방법으로 얻어진 피처 집합을 병합하여 얻어진 결합 피처 집합을 사용하여 기계학습의 예측 성능을 향상시킬 수 있음을 보였다.

2. 피처 선택 방법

2.1 Supervised 피처 선택 방법

Supervised 피처 선택 방법은 어떤 데이터가 참 혹은 거짓으로 분류되는가에 대한 정보를 사용하는 선택 방법이다. 여기서 참 또는 거짓의 값을 갖는 변수를 클래스(class)라 한다. 피처를 F , 클래스를 C 라 할 때 피처와 클래스의 상호정보량(mutual information)은 다음과 같이 정의된다.

$$I(F; C) = \sum_{f \in F} \sum_{c \in C} p(f, c) \log \frac{p(f, c)}{p(f)p(c)} \quad (1)$$

상호정보량의 값이 클수록 해당 피처가 클래스에 대해 많은 정보를 포함하는 것이므로, 이 값의 순위를 보면 피처들의 중요도를 어느 정도 알 수 있다. 하지만 상호정보량 순위만을 보고 피처를 추출하는 방법론은 일반적인 경우에 있어 그다지 좋은 성능을 보여주지 못한다.

본 연구에서는 각 피처의 유효성을 평가하기 위해 중요도(significance)와 종속도(dependence)라는 두 가지 측정 기준을 정의했다(표 1).

알려지지 않은 때에는 서로 독립인 경우가 발생 가능함)

3. 실험

실험에 사용한 데이터는 2003 년도 NIPS(Neural Information Processing Systems) 피쳐 추출에 대한 워크샵(Workshop on Feature Extraction)에서 제시된 다섯 개의 데이터셋(dataset)이다. 이 데이터들은 서로 다른 연구 분야의 데이터 특성을 반영하도록 구성된 것으로서, 피쳐 선택 방법을 개발하고 시험하기에 매우 좋은 대상이다[2].

3.1 실험 데이터의 구성

실험 데이터는 표 2 와 같이 다섯 개의 데이터셋으로 구성되어 있다. 즉 암환자(Arcene), 문자인식(Gisette), 텍스트 분류(Dorothea, Dexter)등 다양한 분야의 데이터를 포함하고 있다. 실험에서는 랜덤(random) 피쳐 벡터와 클래스 변수와의 상호정보량을 기준으로 필터링한 데이터셋을 사용했다.

Name	Train	Validate	Test	Feature
Arcene	100	100	700	10,000
Gisette	6,000	1000	6,500	5,000
Dorothea	800	350	800	100,000
Dexter	300	300	2,000	20,000
Madelon	2,000	600	1,800	500

표 2. 실험 데이터의 구성

(Train, Validate, Test는 해당 데이터 개수, Feature 는 피쳐의 개수임)

3.2 실험 결과

실험 데이터 중 Test 데이터셋의 클래스 정보는 NIPS 워크샵의 대회 평가 목적상 비공개였으므로, Train 데이터셋을 사용하여 학습한 모델을 Validate 셋에 적용하여 테스트 오류(test error)를 측정했다. 각각의 데이터셋에서 추출된 피쳐의 개수는 표 3 과 같다.

실험에 사용한 기계학습 방법론은 Naïve Bayes 와 SVM(Support Vector Machine)이다. 피쳐 집합의 성능 비교를 위해 피쳐와 클래스 변수의 상호정보량 순위를 이용하는 방법(SIMPLE), supervised 피쳐만 이용하는 방법(S), unsupervised 피쳐만 이용하는 방법(US), 병합된 피쳐를 이용하는 방법(MIXED)를 비교 평가했다 (표 4). 표 4 의 진한 회색 부분과 기계학습 오차를 그래프로 정리한 그림 4 를 보면 대부분의 경우 MIXED 피쳐 집합이 가장 좋은 성능을 보임을 확인할 수 있다.

4. 결론

본 논문에서는 기계학습 예측 성능 향상을 위한 간단 하면서 좋은 성능을 보이는 피쳐 선택 방법론을 제안했다. 또 supervised 방법만을 사용한 경우에 비해 결합 피쳐를 사용할 경우 주목할 만한 성능 향상이 있음을 확인했다. 향후 이 방법은 유전자 발현 데이터에서 환자 구별에 중요한 유전자를 찾는 등 타 학문 분야에서 의미가 있는 피쳐를 찾기 위한 방법으로 응용될 수 있다.

감사의 글

본 연구는 과학기술부의 국가지정연구실 사업(NRL) 과 Systems Biology 사업에 의해 지원되었음.

Dataset	Filtered	S	US	MIXED
Arcene	2,882	185	289	417(4)
Gisette	1,958	239	597	766(15)
Dorothea	5,032	827	504	1,017(1)
Dexter	5,951	362	186	467(2)
Madelon	111	39	12	40(8)

표 3. 추출된 피쳐의 개수

(Filtered는 처음 필터링된 피쳐의 수, S, US, MIXED 는 각각 supervised, unsupervised, 병합 피쳐 집합의 개수. 괄호 안의 숫자는 원래 피쳐 대비 %임)

Dataset	SIMPLE		S		US		MIXED	
	NB	SM	NB	SM	NB	SM	NB	SM
Arcene	51	53	31	21	34	35		
Gisette	49	49	11	3	31	39		
Dorothea	13	12	7		17	12		7
Dexter	46	47	17	20	47	56		
madelon	51	51	39		43	41		40

표 4. Validate 데이터셋에 대한 기계학습 오류 (%) (SIMPLE, S, US, MIXED는 각각 순위기반, supervised, unsupervised, 병합 피쳐 집합을, NB와 SM은 각각 Naïve Bayes와 SVM을 나타냄. 진한 회색은 가장 좋은 성능을, 옅은 회색은 두 번째로 좋은 성능을 보인 부분임)

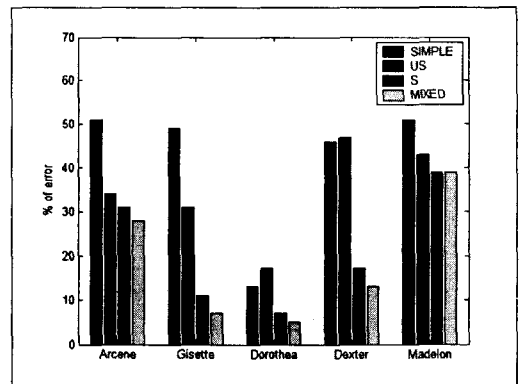


그림 4. 각 피쳐 집합에 대한 기계학습 오류 (%) (Naïve Bayes Classifier)

참고문헌

- [1] Isabelle, G. and Andre, E., An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, vol 3, pp.1157-1182, 2003.
- [2] Isabelle, G., Steve, G., Asa, B.H. and Gideon, D., Benchmark Datasets and Challenge Result Summary, (<http://clopinet.com/isabelle/Projects/NIPS2003/>).
- [3] Richard, O.D., Peter, E.H. and David, G.S., *Pattern Classification*, New York, NY: Wiley-Interscience, 2000.
- [4] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Francisco, CA: Morgan Kaufmann, 1988.
- [5] Spirtes, P., Glymour, C. and Scheines, R., *Causation, Prediction, and Search*, New York, NY: Springer-Verlag, 1993.