정보이론에 기반한 Supervised, Unsupervised 피처 선택 방법론

이상근[0] 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
sklee[0]@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Information-based Supervised and Unsupervised Feature Selection Methods

Sang-Kyun Lee[0] and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

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 정보 이론에 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. 서 론

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 중요한 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.

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

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).

기준	수식
Significance	$I(F_1;C \mid F_2)$
Dependence	$I(F_1;F_2 \mid C)$

표 1. Supervised 피처 선택에서 사용한 두 측정 기준

이 기준에서 사용된 조건부 상호정보량(conditional mutual information)은 다음과 같이 정의된다.

$$I(X;Y \mid Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)\, p(y \mid z)} \quad (2)$$

즉, 이것은 변수 Z가 주어질 경우 변수 X와 Y사이에 남아 있는 상호정보량을 의미한다. 따라서 $I(F_1;C \mid F_2)$는 피처 F2 가 주어졌을 때 피처 F_1이 클래스 C에 대해 갖는 정보량을, $I(F_1;F_2 \mid C)$는 클래스 C의 값이 주어졌을 때 피처 F_1이 서로 다른 피처 F_2에 대해 갖는 정보량을 의미하게 된다.

우리의 관심의 대상이 되는 피처들은 중요도가 높고, 종속도가 낮은 피처들이다. 즉 다른 피처들이 존재할 때 클래스 변수 C 에 대한 정보를 많이 포함하는 동시에, 클래스 변수를 알고 있는 경우 다른 피처들과의 중복도가 적은 피처들을 선택해야 한다. 이러한 동작을 위해 간단한 다목적(multi-objective) 최적화 알고리즘을 고안했으며, 그 개요는 그림 1 과 같다.

Multi-Opt(features)
- Compute significance matrix $I(F_1;C \mid F_2)$
- Compute dependence matrix $I(F_1;F_2 \mid C)$
- Initialize a bucket (size = N)
- For every features in feature pool
  - Move N most significant features to bucket
  - Move N/2 most dependent features from bucket to feature pool
  - Output features in the bucket

그림 1. Supervised 피처 추출을 위한 다목적(multi-objective) 최적화 알고리즘

이 알고리즘은 우선 중요도, 종속도 매트릭스를 계산한 다음, 크기가 N 인 버킷(bucket)을 사용하여 가장 중요도가 높은 N 개의 피처를 추출한다. 그 다음 이 버킷에 있는 피처 중에서 가장 종속도가 높은 N/2 개의 피처를 다시 피처 풀(feature pool)로 돌려보내고, 버킷의 내용을 출력한다. N 이 작을수록 이 알고리즘은 좋은 성능을 보이지만, N 이 너무 작을 경우 알고리즘의 결과가 한 가지 기준으로 편향될 수도 있음으로 주의해야 한다. 본 논문에서는 N=4 의 값을 사용했다.

2.2 Unsupervised 피처 선택 방법

Unsupervised 피처 선택 방법에서는 supervised 방법에서와 달리 클래스 변수에 대한 정보를 사용하지 않는다. 여기서는 모든 피처에 대해 계층적 병합군집화(hierarchical agglomerative clustering)기법을 사용한다. 계층적 병합군집화 방법은 상향식(bottom-up) 군집화 방법의 하나로, 각 데이터 포인트에서 시작하여 가까운 클러스터끼리 병합해 나가는 과정을 거친다. 이 알고리즘의 개요는 그림 2 와 같다[3].

AHC()
- Desired number of clusters = c
- Number of data points = n
- do
  - Find nearest clusters, Ci, Cj
  - Merge Ci & Cj
  - n = n – 1
- until (c=n)
- return c clusters

그림 2. Unsupervised 피처 추출을 위한 계층적 병합군집화 (hierarchical agglomerative clustering) 알고리즘

이 알고리즘은 각 피처간의 거리 정보를 필요로 하는데, 두 피처간의 거리 D 는 두 피처간의 상호정보량의 역수로 정의하였다 (식 (3)).

$$D(F_1,F_2) = I(F_1;F_2)^{-1} \quad (3)$$

즉 두 피처간에 공통적인 정보가 많을수록 두 피처의 거리는 줄어들게 된다.

2.3 결합 피처 선택

마지막으로 supervised, unsupervised 피처 선택 방법으로 추출된 피처 집합을 병합(merge)했다. 이렇게 하는 이유는 supervised 방법만을 사용할 경우 클래스 변수의 존재로 인해 피처간의 상관관계가 잘못 평가될 수 있기 때문이다. 즉 supervised 방법에서는 클래스 변수와의 상관도가 높고 다른 피처에 독립적인 피처를 추출하는데, 이것은 곧 피처와 클래스 변수에 대해 그림 3 과 같은 인과관계를 전제함을 의미한다. 이러한 경우 전제로 사용한 인과관계와 실제 데이터의 확률분포가 일치하지 않는 문제(unfaithfulness)[4, 5]가 발생할 수 있다. 특히 이 경우 각 변수간의 독립성에 대한 평가가 잘못될 수 있는데, 즉 그림 3 에서와 같이 클래스가 알려지지 않을 경우 서로 독립인 (따라서 양자 모두 선택돼야 하는) 피처들이 클래스가 알려진 경우 서로 연관되어 있는 것으로 평가되어 두 피처 중 하나만 선택되는 문제가 발생할 수 있다. Unsupervised 방법으로 선택된 피처를 병합해 사용하면 이 문제를 일정 부분 해결할 수 있을 것으로 예상했으며, 실험을 통해 검증했다.

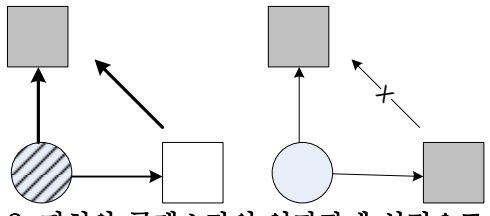

그림 3. 피처와 클래스간의 인과관계 설정으로 인해 피처가 잘못 평가되는 경우의 예
(F1, F2는 피처, C는 클래스 변수, 회색 상자는 선택된 피처, 빗금 표시된 원은 알려진 클래스 변수임. 클래스가 알려진 경우 두 피처가 상관성을 보이지만, 클래스가 알려지지 않은 때에는 서로 독립인 경우가 발생 가능함)

3. 실 험

실험에 사용한 데이터는 2003 년도 NIPS(Neural Information Processing Systems) 피처 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, 피처 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) 피처 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 피처의 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 피처의 개수는 표 3 과 같다.

실험에 사용한 기계학습 방법론은 Naïve Bayes 와 SVM(Support Vector Machine)이다. 피처 집합의 성능 비교를 위해 피처와 클래스 변수의 상호정보량 순위를 이용하는 방법(SIMPLE), supervised 피처만 이용하는 방법(S), unsupervised 피처만 이용하는 방법(US), 병합된 피처를 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 피처 집합이 가장 좋은 성능을 보임을 확인할 수 있다.

4. 결 론

본 논문에서는 기계학습 예측 성능 향상을 위한 간단하면서 좋은 성능을 보이는 피처 선택 방법론을 제안했다. 또 supervised 방법만을 사용한 경우에 비해 결합 피처를 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 피처를 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 피처의 개수
(Filtered는 처음 필터링된 피처의 수, S, US, MIXED 는 각각 supervised, unsupervised, 병합 피처 집합의 개수. 괄호 안의 숫자는 원래 피처 대비 %임)

Feature / Dataset	SIMPLE		S		US		MIXED	
	NB	SM	NB	SM	NB	SM	NB	SM
Arcene	51	53	31	21	34	35	28	14
Gisette	49	49	11	3	31	39	7	3
Dorothea	13	12	7	5	17	12	5	7
Dexter	46	47	17	20	47	56	13	19
madelon	51	51	39	39	43	41	39	40

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

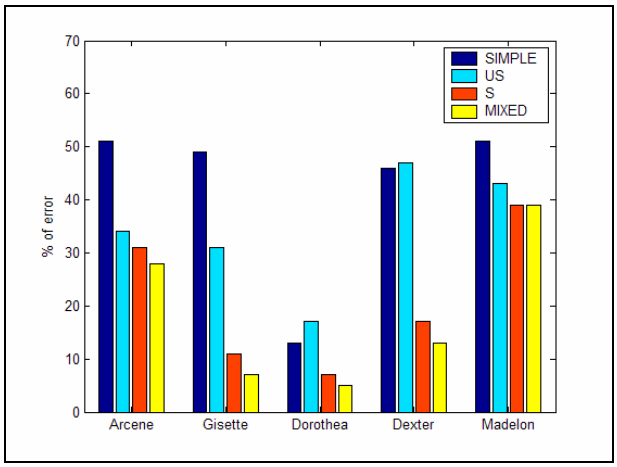

그림 4. 각 피처 집합에 대한 기계학습 오류 (%)
(Naïve Bayes Classifier)

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.