엔트로피를 최대화 하는 실시간 조합 자질 구축 방법

이상우°, 허민오, 장병탁
서울대학교 컴퓨터공학부
{slee, moheo, btzhang}@bi.snu.ac.kr

Online Incremental Associative Feature Construction Methods via Maximizing Entropy

Sang-Woo Lee° , Min-Oh Heo, Byoung-Tak Zhang
School of Computer Science & Engineering, Seoul National University

요 약

데이터를 잘 표현하는 자질 (feature)들을 구축하는 일은 최근 기계 학습에서 활발히 연구되는 핵심적인 문제이다. 본 논문에서는 분류기의 성능을 강화하기 위하여 자질 구축 방법의 일종인 실시간 (online incremental) "조합 자질 구축" (associative feature construction) 방법을 사용한다. 이 때, 특별히 최대 엔트로피 분류기에 본 자질 구축 방법을 적용하여 우수한 분류 성능을 이끌어내었다. 본 논문에서는 제안된 방법을 검증하기 위하여 자연어 처리 문제와 음악 학습 문제를 다룬다. 실험 결과는 데이터를 잘 표현하는 자질들을 뽑아내는 것이 데이터를 설명하는 데에 중요할 뿐 아니라, 실시간 조합 자질 구축 방법이 자질들을 구축하는 좋은 방법임을 보여준다.

1. 서 론

기계 학습 알고리즘을 사용하기 위하여, 전처리 과정으로서 좋은 자질들을 찾아내는 것은 매우 중요하다. 이를 위하여 가장 전통적으로 사용되는 기술은 자질 선택 (feature selection)이다. 자질 선택은 많은 자질 들 중에서 좋은 자질들을 골라내는 것이다. 전통적인 자질 선택 방법은 Wrapper, filter, embedded method 등으로 주로 분류가 가능하며, 각기 많은 연구가 진행되어 왔다 [1].

한편, 자질 구축은 자질 선택과 달리 단순히 주어진 자질들을 사용하는 것이 아닌, 자질들을 혼합하여 새로운 자질들을 만들어 내는 것이다. 커널 기법도 새로운 자질들을 만들어 내는 방법의 한 주요한 예이다. 그러나 본 논문에서 자질 구축은 좀 더 명시적으로 주요한 자질들을 추출해 내는 일을 의미한다. 자질 구축에 대한 배경 이론은 다음과 같다. 데이터를 잘 표현하는 자질들을 구축하는 일은 최근 deep learning 등 기계 학습에서 활발히 연구되는 핵심적인 문제이다. 특별히 deep learning에서 자질 구축이란 restricted Boltzmann machine과 autoencoder의 은닉 변수의 학습을 의미한다 [2]. 최근에는 denoised autoencoder의 은닉 변수의 수를 데이터의 분포에 맞게 실시간으로 조절하는 방법도 제안되었다 [3]. 그러나 이러한 방법은 이산화된 데이터를 다루는 데에 적합할 뿐, 범주화된 데이터를 다루는 데에는 적합하지 않다.

한편, 범주화된 데이터를 다루는 자질 구축 방법에 대한 연구로 하이퍼네트워크 가 있다 [4]. 몇몇 연구들은 하이퍼네트워크를 가지고 진화연산과 유사한 방법을 차용 데이터를 분류하거나, 확률 분포를 학습하는 데 학습하였다 [5, 6]. 또한 이러한 학습 방법은 concept drift를 설명하는 데에 사용되었다 [7]. 본 연구에서는 하이퍼네트워크의 구조학습 방법을 조합 자질을 구축하는 방법으로 이해하고 이 틀을 변형하여 사용한다. 변형된 본 알고리즘은 다음 장에 소개되어 있다. 특별히, 본 논문에서는 자질 구축 방법을 최대 엔트로피 분류기에 사용하였다. 최대 엔트로피 분류기에서 conjunction 형태를 가지는 조합 자질을 사용하는 방법이 여러 논문들에서 소개되어 왔다. 그러나 기존의 연구는 자질의 구축 방법이 휴리스틱한 방법으로 정해지거나 [8], 혹은 실시간으로 이루어지지 않았다 [9]. 실시간으로 데이터를 다루는 일은 빅데이터 문제나 평생 학습과 같은 상황의 학습 문제에서 중요하다. 본 연구는 데이터에 맞게 실시간으로 자질들을 구축하는 데에 관심이 있다.

3. 알고리즘

3.1 조합 자질과 최대 엔트로피 분류기

조합 자질 구축 방법에서 후보가 될 수 있는 조합 자질의 전형적인 예는 다음과 같다.

$$\phi^{(i)}(x,y) = \delta(y,\tilde{y}^{(i)}) \prod_{j \in C^{(i)}} \delta(x_j, \tilde{x}_j^{(i)})$$

이 때, 입력 벡터 $x \in {}^{n}$ 과 그 차원 n에 대하여 $C^{(i)}$ 는 집합 {1, …, n}의 임의의 부분집합이다. 위의 설명을 돕기 위해, 유효한 커널 자질을 예로 들면 다음과 같다.

$$\phi^{(i)}(x,y) = \begin{cases} 1 \ , & if\ x_1 = 1 \& x_3 = 0 \& x_4 = 2 \& y = 1 \\ 0 \ , & otherwise \end{cases}$$

위의 예에서 조합 자질의 $C^{(i)}$는 $\{1, 3, 4\}$이다.

만일 이 조합 자질들을 최대 엔트로피 분류기에 적용하면 다음과 같은 식을 얻을 수 있다.

$$p(y \mid x;\theta) = \frac{\exp(\sum_i w_i \phi^{(i)}(x,y))}{Z}$$

$$s.t.\ Z = \sum_{\tilde{y}} \exp(\sum_i w_i \phi^{(i)}(x,\tilde{y}))$$

이 때, w_i는 커널 함수 $\phi^{(i)}$에 대응되는 가중치이다. 또한 학습식을 다음과 같이 구할 수 있다.

$$\Delta w_i = <\phi^{(i)}(x,y)>_{Data} - <\phi^{(i)}(x,y)>_{P(y|x)}$$

Online 학습을 고려할 때, 다음 추계학적인 학습 방법을 이용할 수 있다.

$$w_i := w_i + \frac{\lambda}{\sqrt{t}} \Delta w_i$$

여기서 λ는 상수, t는 epoch을 의미한다. 추계학적인 방법 방법은 online 학습 상황에서 조건부 우도 값의 하한을 가진다. 구체적으로 batch 방식의 학습과 비교하여 그 오차가 $O(\sqrt{t})$ 임을 보증한다 [10]. 데이터 셋 전체가 학습 시 항상 제공되는 전형적인 기계학습 문제에서, online 학습 방법은 종종 학습 속도를 크게 향상시키며, 좀 더 정확한 확률 분포 학습을 가능하게 하는 batch 학습 방법과 비교해 선택 가능하다. 하지만, 실시간 학습을 필요로 하는 기계학습 문제에서는 online 학습 방법이 필수적이다. 한편, 최대 엔트로피 분류기를 포함한 log linear 모델의 가중치 학습은 convex optimization 문제이다 [11]. 이는 가중치 학습의 결과가 항상 global optima를 보장함을 의미하며, 이러한 성질은 실시간 학습에서 큰 도움이 된다.

조합 자질을 커널 함수에 사용하는 경우는 단일 입력 변수와 클래스 y만의 관계만을 표현하는 고전적인 최대 엔트로피 분류기의 사용과 차이가 있다. 이 차이는 그림 1에 소개되어 있다.

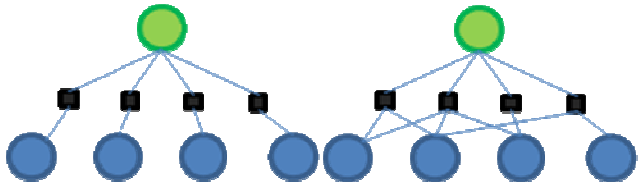

그림 1. (왼쪽) 일반적인 최대 엔트로피 분류기의 factor graph (오른쪽) 조합 자질 구축을 통해 얻어진 커널 함수를 새로운 자질로써 사용하는 최대 엔트로피 분류기의 factor graph. 아래의 동그라미는 입력 벡터 x를, 위의 동그라미는 클래스 y를 의미한다.

```
1  Feature Set: S = empty
2  For n = 1: dataN
3        answer = predict( x^(n) ,S)
4        If answer ~=  y^(n)
5              S = S+ sample( x^(n), y , S);
6        Δw^(n)_{t-k:t}  = learn( x^(n), y ,S);
7        S.w = S.w +  η Δw^(n)_{t-k:t}
8        S = prune(S);
9  End
```

알고리즘 1. 분류기를 위한 실시간 조합 자질 구축 방법

3.2. 실시간 조합 자질 구축 방법

분류기를 위한 실시간 조합 자질 구축 방법은 알고리즘 1에 소개되어 있다. 실시간 조합 자질 구축 방법에서는 자질들의 후보를 실시간 (incremental online)으로 생성 및 제거하고 (5행, 8행), 또한 그 가중치를 online으로 변경한다 (6행, 7행).

한편, 구축 자질들의 후보를 효율적으로 추출하는 데에 (5행) 다양한 방법들이 사용될 수 있다. 예컨대, 자질들의 형태를 실제 데이터의 자질들의 값에서 따오는 방법을 생각해 볼 수 있다. 한편, 커널 자질들을 구성하는 자질들과 클래스간의 관계가 높은 것을 선호하기 위하여, 커널 자질들을 뽑을 때 상호정보량 (mutual information)을 사용할 수 있다. 본 방법에서는 두 가지 아이디어 모두를 사용하여, 분류 성능을 더 개선하였다. 한편, 자질들의 후보를 걸러내기 위하여 (8행) 또한 다양한 방법이 사용될 수 있다. 본 방법에서는 가중치가 낮은 자질들을 제거하는 방법을 사용하였다.

4. 실험 및 결과

제안된 방법의 성능을 보이기 위하여, 본 논문에서는 자연어 처리 문제와 음악 생성 문제를 다루고 그 성능을 보고한다. 첫 번째 문제로 자연어 처리 문제에 해당하는 동일 지시어 문제 (coreference resolution problem)을 다루었다. 동일 지시어 문제란 문서 내에서 등장하는 두 단어가 같은 단어인지 여부를 판단하는 문제이다. 이 문제를 해결하기 위하여 동일 지시어 문제를 분류 문제로 보는 [12]의 패러다임을 차용하였다. 이미, 최대 엔트로피 분류기로 동일 지시어 문제를 해결한 사례가 있다 [13]. 본 논문에서는 자질 구축을 통하여 그 성능을 더욱 높혔다. 실험 자료로는 CONLL 2011의 contest에서 사용된 데이터를 사용하였다. 문서 10개를 학습하고, 또 다른 문서 10개로 이를 검증하여 그 성능을 F-Score로 비교하였다. 그 결과는 표 1과 같다.

NB	SVM*	SVM **	DT	MaxEnt	Suggested Model
0.524	0.428	0.554	0.529	0.537	**0.562**

* SVM, poly kernel
** SVM, RBF kernel, gamma - 0.04
표 1. 동일 지시어 문제에서 알고리즘 간 성능 비교

NB	SVM	DT	Suggested Model
25.89	27.92	23.69	**29.34**

표 2. 노래에서 다음 음을 맞추는 문제에서 알고리즘 간 성능 비교

표에서, NB는 naïve Bayes, SVM은 support vector machine, DT는 decision tree, MaxEnt는 최대 엔트로피 분류기를 의미한다. 한편, Suggested model은 앞에서 설명한 바와 같이 최대 엔트로피 분류기에 실시간 자질 구축 방법을 적용한 모델이다. 이러한 실험 결과는 제안된 방법이 기존의 최대 엔트로피 분류기뿐 아니라 다른 전통적인 분류기들보다 좋은 성능을 가짐을 보여준다.

두 번째 과제는 노래의 다음 음이 무엇인 지를 맞추는 음 예측 문제이다. 이 문제는 실제 노래에서 현재 음까지의 음(Pitch)를 듣고 다음 음이 무엇인 지 맞추는 문제이다. Data는 비틀즈의 노래 40곡의 midi 이며, 곡 단위로 잘라 10-cross validation을 수행하고 그 성능을 Accuracy로 비교하였다. 그 결과는 표 2와 같다. 이를 통해 상이한 과제에서 제안된 모델이 좋은 성능을 보임을 확인할 수 있다.

5. 결론

본 논문에서는 실시간 조합 자질 구축 방법이 데이터의 조건부 확률을 학습하는 데에 효과적임을 논증하였다. 특별히 최대 엔트로피 분류기의 자질을 학습하는 데에 실시간 조합 자질 구축 방법이 효과적임을 보였다. 그러나 실시간 조합 자질 구축 방법은 전통적인 기계학습 분류 문제 및 예측 문제 보다는, 실시간으로 학습을 수행해야 하는 빅데이터 문제나 평생 학습과 같은 주제에서 더 중요하게 사용될 수 있다. 추후에는 위에서 언급된 주제에 해당되는 문제를 찾아 본 실시간 조합 자질 구축 방법을 시험해보고자 한다.

본 논문에서는 조합 자질 구축 방법을 분류 성능을 강화하는 데에 사용했다. 하지만, 생성 모델 (Generative Model)이 데이터의 확률 분포를 추론하는 데에도 같은 방법이 사용될 수 있다 [4, 6, 7]. 조합 자질 구축 방법을 통해 얻어진 자질들은 SVM과 같은 모델과 달리 중요한 자질들이 명시적으로 드러나게 되며, 따라서 데이터를 요약하는 데에도 사용될 수 있다 [5-7]. 마지막으로 분류기를 학습하기 위해 구축된 자질들은 또한 다른 분류기의 자질로서 효과적으로 사용될 수 있다 [5].

감사의 글

이 논문은 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734-Videome, NRF-2013M3B5A2035921-HyperIntelligence), 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원 지원(KEIT-10035348-mLife, KEIT-10044009)을 일부 받았음.

참고 문헌

[1] I. Guyon, A. Elisseeff, “An Introduction to Variables and Feature Selection”, *JMLR*, 2003.
[2] Y. Bengio, “Learning Deep Architectures for AI”, *Foundations and Trends in Machine Learning,* 2009.
[3] G. Zhou, K. Sohn, H. Lee, “Online Incremental Feature Learning with Denoising Autoencoder” *AISTATS*, 2012.
[4] B. -T. Zhang, “Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory”, *IEEE Computational Intelligence Magazine*, 2008.
[5] E. -S. Kim, J. -W. Ha, B. -T. Zhang, “Mutual information-based evolution of hypernetworks for brain data analysis”, *CEC,* 2011.
[6] Ha, “Text-to-image retrieval based on incremental association via multimodal hypernetworks“, *IEEE SMC,* 2012.
[7] B. -T. Zhang, “Sparse population code models of word learning in concept drift”, *Cogsci*, 2012.
[8] K. Toutanova, D. Klein, C. D. Manning, Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”, *NAACL*, 2003.
[9] S. -B. Park, B. -T. Zhang, “A boosted maximum entropy model for learning text chunking”, *ICML,* 2002.
[10] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, *Association for the Advancement of Artificial Intelligence*, 2003.
[11] D. Koller, N. Friedman, “Probabilistic graphical models: principles and techniques”, 2009.
[12] W. M. Soon, H. T. Ng, D. C. Y. Lim, “A Machine Learning Approach to Coreference Resolution of Noun Phrases”, *Computational Linguistics*, 2001.
[13] S. P. Ponzetto, M. Strube, “Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution”, *HLT-NAACL*, 2006.