특징 선택 방법들의 성능 비교 분석에 대한 연구

윤철민, 양지훈
서강대학교 컴퓨터학과
e-mail : toro1983@gmail.com

An Experimental Study on Feature Subset Selection Methods

Chulmin Yun, Jihoon Yang
Dept. of Computer Science, Sogang University

요 약

패턴 인식의 성능 향상을 위해 효율적인 특징 선택을 해주는 것은 중요한 과정 중 하나이다. 본 연구에서는 최근에 제시되었던 특징 선택 방법들 중 대표적인 4 가지의 방법을 선택하여 성능을 비교 분석하였다. 데이터의 특징을 줄여주는 기능과 적은 수의 특징으로 더 좋은 패턴 인식 성능을 보여주는 기능의 수행 능력을 중심으로 평가하였으며, 각기 다른 형태의 데이터에 대해 각 방법들이 보이는 성능도 관찰하였다. 이를 토대로 각 방법의 장단점과 특징에 대해 생각해 본다.

1. 서론

특징 선택 (Feature Selection) 은 패턴 인식 문제에서 중요한 이슈 중의 하나로, 데이터의 차원을 줄여 패턴 인식기의 수행 시간을 단축시켜주고 수행 성능을 더 좋게 하는 효과를 가져온다.
오랜 시간에 걸쳐 다양한 특징 선택 방법들이 개발·논의되어 왔으며, 현재도 꾸준히 새로운 방법들이 제안되고 있다. 본 연구에서는 최근에 제시되었던 특징 선택 방법들 중 몇 가지를 선택하여 실험을 통해 각 방법들의 성능을 비교 분석해 보기로 한다.

2. 특징 선택 방법

비교 분석을 위해 채택한 특징 선택 방법은 총 4 가지이며, 각각의 방법에 대해 간략하게 소개하면 다음과 같다.

2.1 mRMR (H. Peng, 2005)

이 방법은 특징들 간의 상호 정보 (mutual information) 을 기반으로 한 것으로, 상호 정보는 두 개의 대상이 얼마나 서로 의존적인지를 알려준다. 이 방법에서는 특징과 클래스와의 상호 정보량을 계산하여 클래스에 대한 연관성(relevance) 의 척도로 삼고, 그 특징과 다른 특징들과의 상호 정보량을 계산하여 잉여성(redundancy) 의 척도로 삼는다. 이 두 값의 차를 특징의 중요도로 삼게 되는데, 즉 두 값의 차가 클수록 잉여성은 적고(minimal-redundancy) 연관성은 커지는(maximal-relevance) 특징이 되어 패턴 인식에 있어 중요도가 큰 특징이라고 볼 수 있는 것이다.

2.2 CMIM (Fleuret, 2004)

이 방법은 조건부 상호 정보 측정 (conditional mutual information measurement) 을 특징 선택의 기준으로 한다. 이를 이용해 이미 선택된 특징을 전제로 새롭게 선택하려는 특징이 클래스와 얼마나 많은 상호 정보량을 가지게 되는가를 계산하여 특징 선택의 기준으로 삼는다. 본 연구에서 사용한 이 방법은 계산의 복잡성으로 인해 이진 형태(Binary Type)를 가지는 특징과 클래스에 대해서만 적용되었다.

2.3 I-Relief (Yijun Sun, 2006)

Relief (Kira, Rendell, 1992) 방법은 특징 공간 (feature space) 에서 이웃하는 샘플들의 클래스 정보와 거리를 이용한다. 임의의 샘플을 선택하여 그와 같은 클래스

※ 본 연구는 한국과학재단 특정기초연구 (R01-2004-000-10689-0) 지원으로 수행되었음

를 가지는 샘플과의 거리와 다른 클래스를 가지는 샘플과의 거리를 이용하여 계산한 값을 특징에 대한 가중치(weight)로 삼는다. Relief 방법은 계속하여 개선되어 왔으며, 본 연구에서 사용한 방법은 Relief 의 치우침(bias)을 보완하기 위한 Iterative-Relief 방법이다.

2.4 Genetic Algorithm

유전자 알고리즘은 잘 알려진 진화 적응형 알고리즘의 하나로, 특징 선택 분야에서도 유용하게 사용되고 있는 방법이다. 특징 선택에서는 데이터 안의 모든 특징들이 특징의 수만큼의 길이를 가지는 비트 문자열로 나타내어지며 문자열의 각 자리에 해당하는 특징이 선택되었을 때는 1 로, 선택되지 않았을 때는 0 으로 나타낸다. 이 비트 문자열을 유전자 알고리즘의 인자(gene)로 삼아 일반적인 유전자 알고리즘을 적용하여 가장 높은 적합성(fitness)을 보이는 특징들의 조합을 찾아낸다. 적합도는 기계학습 알고리즘을 사용한다.

3. 실험

비교 분석을 위해 총 다섯 개의 데이터를 사용하였다. 모든 데이터는 UCI Machine Learning Repository 에 있는 실제 데이터들이다. 모든 데이터들이 연속적인 값을 가지는 데이터들이기 때문에 값을 이산화하여 사용하였고, CMIM 과 다른 방법들과의 비교를 위해 추가로 세 개의 데이터 (Arrhythmia, Sonar, Ionosphere)에 대해선 평균값을 기준으로 값을 이진화 한 데이터들을 만드는 전처리 과정을 거쳤다.

<표 1> 실험에 사용한 데이터에 대한 요약

	특징 수	샘플 수	클래스 수
Arrhythmia	277	452	16
Image segmentation	19	2310	7
Waveform	21	5000	3
Sonar	60	208	2
Ionosphere	34	351	2

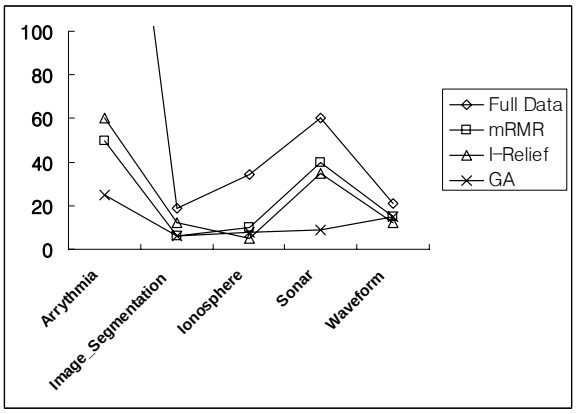
각 방법을 통해 선택된 특징들의 성능 평가를 위해 JAVA 기반의 패턴 인식 도구인 WEKA 를 사용하였다. 성능 평가에는 베이즈 결정 이론을 기반으로 한 Naïve-Bayes 와 결정 트리 기반의 C4.5 의 두 패턴 인식 알고리즘이 사용되었으며, 두 알고리즘 모두 10 겹 교차 검증 (10-fold Cross-Validation) 을 이용하여 분류의 정확도를 측정하였다.

4. 결과 분석

먼저, 특징 선택을 하지 않은 상태인 모든 특징이 포함된 상태에서 각 데이터의 정확도를 측정하였다. 그리고 각 방법을 통해 특징을 선택하고 선택된 특징들만으로 구성한 데이터에 대해 정확도를 측정하여 비교하였다.

4.1 선택된 특징의 수 비교

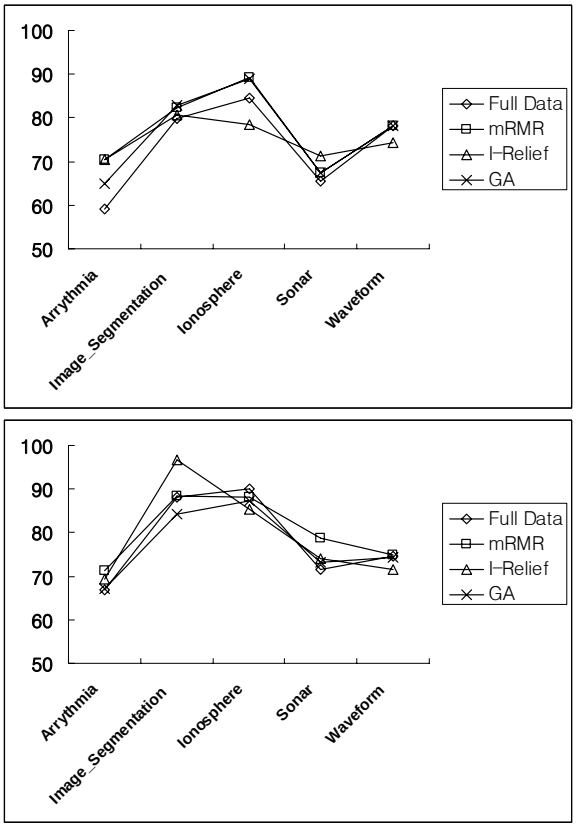
먼저 CMIM 을 제외한 세 가지 방법을 이산화한 데이터에 적용하여 선택된 특징의 수를 비교해 보면, 각각의 방법 모두 전체 데이터의 특징 수 보다 더 적은 수의 특징을 선택해 주는 것을 확인할 수 있다. 각



(그림 1) 선택된 특징의 수 비교

방법 모두 데이터의 차원을 줄여 주는 역할은 문제없이 잘 수행해 주고 있다는 것을 보여주는 결과이며, 특히 유전자 알고리즘의 경우 다른 방법들에 비해 대부분의 데이터에서 더 적은 수의 특징을 선택해 주는 것을 확인할 수 있다.

4.2 정확도의 비교



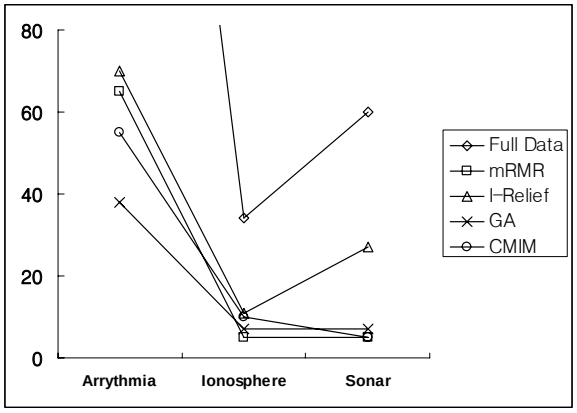
(그림 2) 선택된 특징으로 구성된 데이터들의 정확도 (위 : Naïve-Bayes, 아래 : C4.5)

특징 선택 방법들을 통해 선택된 특징들로 구성한 데

이터들의 정확도를 살펴보면, 대부분의 방법들이 대부분의 데이터에서 모든 특징을 포함하였을 때보다 특징을 선택하여 골라냈을 때 더 좋은 정확도를 보여주는 것을 볼 수 있다. 특히 mRMR 방법은 거의 모든 데이터에서 다른 방법들보다 더 좋은 성능을 보여준다. 그러나 I-Relief 방법은 몇몇 데이터에서는 오히려 모든 특징을 포함하였을 때의 데이터보다 더 낮은 정확도를 보여주는 데, 이는 특징공간에서의 이웃하는 샘플들 간의 거리를 기반으로 하는 Relief 방법의 특성상 이산화된 데이터에서는 어느 정도의 치우침 현상이 발생하기 때문인 것으로 생각된다.

4.3 이진 데이터에 대한 성능 비교

CMIM 방법을 포함한 비교를 위해 세 가지의 이진화 한 데이터로 위와 동일한 실험을 행하였다. 역시 모든 방법들이 더 적은 수의 특징을 선택해주었다. 반면 정확도 면에서는 mRMR 과 CMIM 만이 좋은 정확도를 보여주었으며 유전자 알고리즘은 Ionosphere 데이터에 대해서는 좋은 정확도 향상을 보였지만 그 밖의 데이터에 대해서는 위의 두 방법들에 비해 좋은 성능을 보여주지 못했다. I-Relief 방법은 다른 방법들과 크게 비교될 만큼의 정확도 저하를 보여 주었는데, 역시 이산화 된 데이터에서 나타났던 치우침 현상이 이진화 된 데이터에서는 더 심해지는 것으로 보인다.



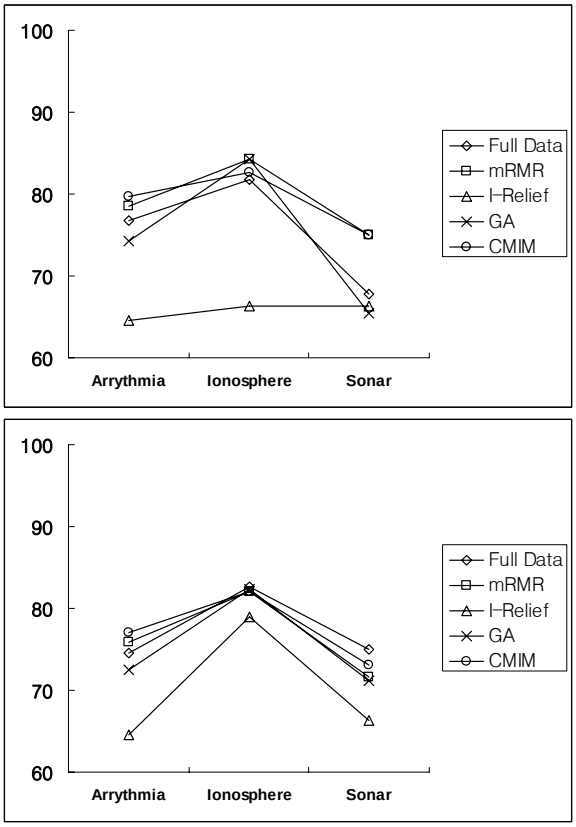
(그림 3) 이진화 한 데이터에서의 선택된 특징의 수

5. 결론

지금까지 최근에 제시된 특징 선택 방법 4 가지를 실험을 통해 비교해 보았다. 거의 대부분의 경우 특징 선택을 통해 특징의 수를 줄이고 정확도의 향상을 가져오는 것을 확인할 수 있었다.

mRMR 의 경우 가장 좋은 성능을 보였다고 할 수 있는데, 특징 선택 방법의 중요한 역할인 특징 수 줄이기와 성능 향상을 함께 가져오는 데 있어 대부분의 데이터에서 다른 방법들보다 대체로 좋은 결과를 보여주었다. CMIM 도 무난한 성능을 보였으나 이진화된 데이터에서밖에 사용할 수 없다는 것이 상대적인 결함으로 보인다.

I-Relief 의 경우 상대적으로 좋지 않은 성능을 보였는데, 이는 위에서도 밝힌 바와 같이 이산화, 이진화된 데이터에서는 치우침 현상이 생기게 되어 특징의 가중치가 고르게 계산되지 않기 때문인 것으로 보이며 따라서 연속형 데이터에서는 더 좋은 성능을 보여줄 것으로 기대된다.



(그림 4) 이진화 한 데이터에서의 선택된 특징으로 구성된 데이터들의 정확도
(위 : Naïve-Bayes, 아래 : C4.5)

참고문헌

[1] Hanchuan Peng, Fuhui Long and Chris Ding "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, 2005.

[2] Francois Fleuret "Fast Binary Feature Selection with Conditional Mutual Information" Journal of Machine Learning Research 5 1531-1555, 2004.

[3] Yijun Sun and Jian Li "Iterative RELIEF for Feature Weighting" Proceedings of the 23rd International Conference on Machine Learning, 2006.

[4] K.Kira and Larry A. Rendell "A Practical approach to feature selection" Proceedings of the 9th International Workshop on Machine Learning, 1992.

[5] UCI ML Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html

[6] WEKA, http://www.cs.waikato.ac.nz/~ml/index.html

[7] Avrim L. Blum and Pat Langley "Selection of Relevant Features and Examples in Machine Learning" ACM Artificial Intelligence Archive, Vol. 97, Issue 1-2, 245-

271, 1997.
[8] Isabelle Guyon and Andre Elisseeff "An Introduction to Variable and Feature Selection" Journal of Machine Learning Research 3 , 1157-1182, 2003.