# 특징 래핑을 통한 숫자형 특징과 범주형 특징이 혼합된 데이터의 클래스 분류 성능 향상 기법

## (Improving Classification Performance for Data with Numeric and Categorical Attributes Using Feature Wrapping)

이 재 성 † 김 대 원 ††
(Jaesung Lee) (Daewon Kim)

**요 약** 본 논문에서는 혼합형 데이터에 대한 특징 선별 기법의 효율성을 비교하기 위해 특징 필터링과 특징 래핑을 통한 특징 선별 후, 클래스 분류 성능을 측정하였다. 혼합형 데이터는 숫자형 특징과 범주형 특징이 함께 혼합되어 있으므로, 숫자형 특징을 범주형 특징으로 이산화를 하여 단일형 데이터로 변환한 뒤 특징 선별 기법 등을 적용할 수 있다. 본 연구에서는 혼합형 데이터를 전처리하여 단일형 데이터로 변환하고, 널리 활용되는 특징 필터링 기법과 특징 래핑 기법을 통해 클래스 분류 성능을 높일 수 있는 특징 집합을 선별하였다. 선별된 특징 집합을 통한 클래스 분류 성능을 비교한 결과, 특징 필터링에 비해 특징 래핑을 통해 선별한 특징 집합을 활용하여 클래스 분류를 하였을 때 분류 정확도가 높은 것을 확인할 수 있었다.

**키워드** : 클래스 분류, 혼합형 데이터, 특징 선별

***Abstract*** In this letter, we evaluate the classification performance of mixed numeric and categorical data for comparing the efficiency of feature filtering and feature wrapping. Because the mixed data is composed of numeric and categorical features, the feature selection method was applied to data set after discretizing the numeric features in the given data set. In this study, we choose the feature subset for improving the classification performance of the data set after preprocessing. The experimental result of comparing the classification performance show that the feature wrapping method is more reliable than feature filtering method in the aspect of classification accuracy.

**Key words** : Classification, Mixed-type data, Feature wrapping

## 1. 서 론

임상의학이나 비즈니스 분야 등 기계 학습이 적용될 수 있는 분야가 넓어지면서, 키, 성별, 질병의 종류 등 혼합형 데이터에 대한 대응 능력과 높은 클래스 분류 정확도가 크게 요구받고 있다[1,2].

기존 제안된 데이터 마이닝 기법들은 단일형 데이터에 특화되게 설계되어 있어 혼합형 데이터에 그대로 적용하기가 어렵다. 따라서 혼합형 데이터에 대해 데이터 마이닝 기법을 적용하기 위해 단일형 데이터로 전처리를 하는 방법이 널리 활용되어 왔다.

한편, 데이터에 대한 클래스 분류 성능을 높이기 위해 단일형 데이터를 대상으로 특징 선별 기법이 널리 활용되어 왔다. 특징 선별 기법에는 특징들에 대한 랭킹 평가를 통해 개별 특징을 평가하고, 순위에 따라 특징을 선별하는 특징 필터링(Feature Filtering)과 클래스 분류 성능을 높일 수 있는 특징 집합을 점진적으로 구성해나가는 특징 래핑(Feature Wrapping)으로 나누어진다.

혼합형 데이터에 대해 특징 선별 기법을 적용하기 위해서는 다음과 같은 문제를 해결하여야 한다. 첫째, 특징

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2009-0068718).

† 학생회원 : 중앙대학교 컴퓨터공학과
jslee.cau@gmail.com
†† 정 회 원 : 중앙대학교 컴퓨터공학과 교수
dwkim@cau.ac.kr

논문접수 : 2009년 8월 20일
심사완료 : 2009년 9월 30일

필터링을 적용할 경우 앞서 언급하였듯이 필터링 기법 자체가 범주형이나 수치형 특징 중 한 가지 형태에 특화되어 있어 혼합형 데이터에 그대로 적용하기 어렵다. 둘째, 특징 래핑을 통한 특징 선별의 경우 특징 선별을 위한 학습 알고리즘에 따라 혼합형 데이터에 적합한 유사도 측정 기법을 활용해야 하는 제한이 따른다. 그러나 특징 선별 기법은 클래스 분류에 도움이 되지 않는 특징을 제거함으로써 클래스 분류 성능을 높이거나, 클래스 분류 성능을 유지하면서 전체 데이터에 비해 적은 특징 집합을 얻을 수 있다는 점에서 매우 중요하다.

본 연구에서는 특징 선별 기법의 두 가지 큰 분류인 특징 필터링과 특징 래핑의 효율성을 일관적으로 비교하기 위해 혼합형 데이터에 대한 전처리를 통해 숫자형 특징을 범주형 특징으로 전환하여, 단일형 데이터로 변경한 뒤 특징 선별 기법을 적용하였으며, 이를 통해 클래스 분류 성능을 높일 수 있는 특징들을 선별하였다. 또한, 선별된 특징의 클래스 분류 성능을 비교하여 혼합형 데이터에 대해 어떠한 특징 선별 기법이 효율적인지를 보인다.

본 논문이 가지는 혼합형 데이터에 대한 연구 기여는 혼합형 데이터의 클래스 분류 성능을 높이기 위해 특징 필터링과 특징 래핑 중 어떠한 기법을 적용하는 것이 적합한지 밝히는데 있다.

## 2. 비교 알고리즘

### 2.1 특징 래핑

본 논문에서는 특징 집합에 대한 평가 함수로 정확도를 활용하여 선택된 특징 집합의 우수함을 평가하면서, 특징의 조합을 찾아나가는 특징 래핑 기법을 적용하였다. 특징 래핑 알고리즘은 크게 2가지 알고리즘으로 구성되어 있는데[3], 특징 집합을 구성하기 위한 탐색 알고리즘과 찾아낸 특징 집합을 평가하기 위한 학습 알고리즘이다.

탐색 알고리즘은 효과적인 특징 집합을 찾게 하기 때문에 매우 중요하다. 비록, Exhaustive Search를 통해 최적 특징 집합을 찾을 수 있다하더라도, 특징의 개수가 많은 데이터에 대해서는 적합하다고 볼 수 없다. 시간, 계산상의 비용을 줄이기 위해 본 논문에서는 휴리스틱 탐색 알고리즘으로 널리 활용되고 있는 Sequential Forward Selection(SFS)을 활용하였다.

SFS는 순차적으로 특징을 하나씩 추가하면서 특징 집합을 구성해나간다. 임의의 데이터에 대해서 SFS를 통해 이루어지는 특징 선별 과정은 표 1에서 보이고 있다.

1. 초기화: 특징이 없는 공집합 $S$에서 시작한다.
2. **Step 1**: 주어진 8개의 특징 중 무작위로 하나의 특징($f_5$)를 선택하여, $f_5$를 특징 집합 $S$에 포함시킨다.

표 1 임의의 데이터에 대한 특징 래핑의 프로시저

| Step | 특징 집합 ($S$) | 정확도 | 동작 |
|---|---|---|---|
| 초기 | { } | 00.0% | |
| 1 | { $f_5$ } | 50.0% | $f_5$ 유지 |
| 2 | { $f_5,f_4$ } | 66.7% | $f_4$ 유지 |
| 3 | { $f_5,f_4,f_1$ } | 66.7% | $f_1$ 제거 |
| 4 | { $f_5,f_4,f_3$ } | 83.3% | $f_3$ 유지 |
| 5 | { $f_5,f_4,f_3,f_6$ } | 50.0% | $f_6$ 제거 |
| 6 | { $f_5,f_4,f_3,f_7$ } | 83.3% | $f_7$ 제거 |
| 7 | { $f_5,f_4,f_3,f_8$ } | 100.0% | $f_8$ 유지 |
| 8 | { $f_5,f_4,f_3,f_8,f_2$ } | 100.0% | $f_2$ 제거 |
| 9 | { $f_5,f_4,f_3,f_8$ } | 100.0% | 중지 |

   $S$에 대해 클래스 분류 정확도를 측정한다. 측정 결과 공집합에 대해서 클래스 분류 정확도가 상승(0% → 50%)하였으므로, $f_5$를 $S$에 남겨둔다.
3. Step 2: $f_5$와 마찬가지로 $f_4$ 역시 $S$에 포함한다.
4. Step 3: $f_1$을 $S$에 포함하였을 때 클래스 분류 정확도가 감소하므로, $f_1$을 제거한다.
5. Step 4: $f_3$을 $S$에 추가하고, 정확도를 평가한 뒤 $S$에 남겨둔다.
6. Step 5, 6, 7, 8: $f_8$을 $S$에 남겨두고 $f_6$,$f_7$,$f_2$를 포함하지 않는다.
7. 종료: 모든 특징에 대해서 평가가 완료되었다. 또한 특징 집합 $S$={$f_5$,$f_4$,$f_3$,$f_8$}일 때, 최선의 정확도를 보였으므로 특징 집합 $S$를 선택한다.

특징 래핑에 의해 선택된 특징 집합에 대한 평가를 위해 특징 집합에 대한 평가 함수가 필요하다. 본 논문에서는 클래스 분류 정확도를 특징 집합에 대한 평가 함수로 활용하였으며, 정확도를 측정하기 위해 Value Difference Metric을 유사도 측정 기법으로 활용하는 최우도 분류(Nearest Neighbor)를 활용하였다(NN+VDM)[4].

### 2.2 특징 필터링

특징 필터링은 개별 특징에 대한 평가를 통해 특징 랭킹을 하고, 상위에 랭크된 일정 개수의 특징을 선별한다. 개별 특징에 대한 평가를 통해 랭킹을 결정하기 때문에 특징 간 연관관계를 고려할 수 없다는 단점이 있다. 그러나 각 특징마다 한 번의 평가를 통해 랭킹이 결정되므로, 특징 래핑에 비해 빠르게 특징을 선별할 수 있으며, 클래스 분류기에 관계없이 일반적으로 클래스 분류 성능이 좋은 특징을 선별할 수 있다는 장점이 있다.

특징 필터링으로써 널리 활용되는 기법에는 Information Gain, Twoing Rule, Max Minority, Gini Index 등이 있다[5]. 특징 필터링 기법에 따라 이진 특징에 특화된 경우가 있어, 본 논문에서는 2개 이상의 범주를 가

표 2 특징 선별 기법과 클래스 분류기에 의한 각 데이터의 클래스 분류 성능 비교. 클래스 분류 성능은 100번 반복하여 얻은 정확도에 대해 평균을 적용하여 측정함

| 클래스 분류기 | | | NN+HD | | NN+VDM | |
|---|---|---|---|---|---|---|
| 데이터 | | | Hepatitis | Credit Approval | Hepatitis | Credit Approval |
| 특징 선별 기법 미적용 | | | 78.3% | 80.5% | 79.3% | 83.6% |
| 특징 선별 기법 적용 | 특징 필터링 | Information Gain | 78.2% | 84.4% | 77.5% | 84.4% |
| | | Twoing Rule | 79.0% | 84.5% | 78.3% | 84.5% |
| | | Max Minority | 78.7% | 82.2% | 77.9% | 80.8% |
| | | Gini Index | 79.0% | 84.5% | 78.3% | 84.5% |
| | 특징 래핑 | | 79.3% | 85.0% | 79.3% | 84.5% |

지고 있는 특징을 이진 특징으로 변경하여 특징 필터링 기법을 적용하였다. 특징 필터링 기법 중의 일부는 이진화된 특징의 성능을 평가할 수 있도록 설계되어 있다. 본 연구에서 활용한 특징 필터링 기법들은 일관성을 유지하기 위하여 범주형으로 전처리된 모든 특징을 이진화한 뒤, 위의 특징 필터링 기법을 적용하였다. 임의의 $m$번째 특징 $f_m$에 대한 각각의 특징 필터링 기법의 평가 방법은 다음의 각 수식을 따른다.

1. Information Gain $ig(f_m)$

$$ig(f_m)=\sum_{i=1}^{k}\left(\frac{l_i}{n}log\frac{l_i}{n_l}+\frac{r_i}{n}log\frac{r_i}{n_r}\right)-\sum_{i=1}^{k}\left(\frac{l_i+r_i}{n}\right)\log\left(\frac{l_i+r_i}{n}\right) \quad (1)$$

2. Twoing Rule $tr(f_m)$

$$tr(f_m)=\frac{n_l n_r}{n^2}\left(\sum_{i=1}^{k}\left|\frac{l_i}{n_l}-\frac{r_i}{n_r}\right|\right)^2 \quad (2)$$

3. Max Minority $mm(f_m)$

$$mm(f_m)=\max\left(\sum_{i=1}^{k}l_i-\max(l_i),\sum_{i=1}^{k}r_i-\max(r_i)\right) \quad (3)$$

4. Gini Index $gi(f_m)$

$$gi(f_m)=\frac{n_l}{n}\left(1-\sum_{i=1}^{k}\left(\frac{l_i}{n_l}\right)^2\right)+\frac{n_r}{n}\left(1-\sum_{i=1}^{k}\left(\frac{r_i}{n_r}\right)^2\right) \quad (4)$$

이 때, $n$은 $f_m$에 포함된 패턴의 개수를 의미하고, $k$는 클래스의 개수이며, $n_l$은 특징 $f_m$에서 0으로 표현된 패턴의 개수, $n_r$은 1로 표현된 패턴의 개수이다. 또, $l_i$는 클래스 $i$에 속하는 패턴 중 0으로 표현된 패턴의 개수이고, $r_i$는 클래스 $i$에 속하는 패턴 중 1로 표현된 패턴의 개수를 의미한다.

## 3. 실험 결과

특징 필터링과 특징 래핑을 통한 특징 선별의 효율성을 비교하기 위해 본 논문에서는 크게 2개의 실험을 디자인하였다. 먼저 혼합형 데이터에 포함된 모든 특징들을 범주형 특징으로 변환하여 단일형 데이터로 변환을 하였다. 그 후, 실험 데이터 중 20%에 해당하는 데이터를 별도로 분리하여 테스트 데이터로 활용하였고, 나머지 학습 데이터에 대해 특징 필터링과 특징 래핑을 통해 특징을 선별하였다. 특징 필터링 기법을 적용하여 전체 특징 중 상위 30%의 특징을 선별하였으며, 선별된 특징 집합을 활용하여 별도로 분리한 테스트 데이터에 대한 NN+VDM과 NN+HD의 성능을 측정하였다. 실험 결과의 신뢰도를 높이기 위하여 위의 실험 과정을 100번 반복하고, 이에 대한 실험 결과를 비교하였다.

실험 데이터로는 Hepatitis, Credit Approval로 알려진 2개의 데이터에 대해 특징을 선별하고, 클래스 분류 성능을 비교하였다[6]. Hepatitis 데이터는 150개의 패턴을 포함하고 있으며, 6개의 숫자형 특징, 13개의 범주형 특징으로 구성되어 있다. 또한, Credit Approval 데이터는 690개의 패턴을 포함하고 있으며, 각각의 패턴은 6개의 숫자형 특징과 9개의 범주형 특징으로 이루어져 있다.

표 2는 Credit Approval과 Hepatitis 데이터에 대한 NN+HD와 NN+VDM의 클래스 분류 평균 성능을 보이고 있다. Credit Approval 데이터의 경우 특징 선별 기법을 적용하지 않았을 때, 80.5% 정도의 성능을 보이고 있으나 특징 필터링을 적용하여 특징을 선별하였을 때 2-4%정도의 성능 향상이 있는 것을 확인할 수 있다. 반면에 특징 래핑을 통해 특징을 선별하였을 때에는 4.5%의 성능 향상이 있는 것을 알 수 있다. Hepatitis 데이터의 경우 특징 필터링에 비해 특징 래핑을 통해 특징을 선별하였을 때, 2% 정도의 성능 차이가 나는 것을 확인할 수 있다.

특징 필터링 실험의 경우, 필터링 기법을 적용하지 않았을 때의 성능보다 오히려 좋지 않거나 비슷한 성능을 내는 경우가 있음을 확인할 수 있다. 그러나 특징 필터링을 통해 선별된 특징의 개수가 전체 특징의 개수 중 30%에 불과한 것을 감안하면, 적은 개수의 특징으로도 비슷한 성능을 보이는 것으로 분석할 수 있다. 따라서

특징 필터링 역시 클래스 분류 성능을 높이는데 효과적임을 알 수 있다. 그러나 특징 필터링 기법은 전체적으로 특징 래핑에 비해 성능이 뒤떨어지는 것을 확인할 수 있다. 이는 특징 필터링 기법을 통해 선별된 특징들은 개별적으로 평가가 이루어지기 때문에 특징 사이의 연관관계를 고려할 수 없는 반면에, 특징 래핑 기법을 통해 특징을 선별할 때에는 특징 집합의 성능을 평가하기 때문에 특징의 조합으로 인한 클래스 분류 성능 향상을 기대할 수 있다.

클래스 분류 실험을 통해 보았을 때, 클래스 분류기와 특징 선별 기법에 대해 다소간 차이가 있으나 대체로 특징 선별을 하였을 때 성능의 향상이 있으며, 특히 특징 래핑을 적용하였을 때 항상 최선의 성능이 측정됨을 알 수 있다.

## 4. 결 론

혼합형 데이터에 대한 클래스 분류 성능을 높이기 위해서는 특징 선별 기법이 널리 활용된다. 그러나 특징 선별 기법이 특징 필터링과 특징 래핑으로 서로 다른 성질의 기법이 제안되었고, 특징 필터링과 특징 래핑 중 어떠한 기법이 더욱 효과적인지는 거의 알려져 있지 않았다.

본 논문에서는 혼합형 데이터에 대한 특징 필터링과 특징 래핑의 적용을 통해 클래스 분류 성능을 비교함으로써 어떠한 기법이 더욱 효과적인지 알아볼 수 있었다.

## 참 고 문 헌


[1] K. Cios and G. W. Moore, "Uniqueness of Medical Data Mining," *Artificial Intelligence in Medicine journal*, vol.26, no.1, pp.1-24, Sep, 2002.

[2] E. Tuv, A. Borisov and K. Torkkola, "Unsupervised learning with mixed numeric and nominal data," *IEEE Transactions on Knowledge and Data Engineering*, vol.14, no.4, pp.673-690 2002.

[3] Z. Sun, G. Bebis, and R. Miller, "Object detection using feature subset selection," *Pattern recognition*, vol.37, no.11, pp.2165-2176, Nov, 2004.

[4] D. R. Wilson, and T. R. Martinez, "Improved Heterogeneous Distance Functions," *Journal of Artificial Intelligence Research*, vol.6, no.1, pp.1-34, Jun, 1997.

[5] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and Simon Kasif, "Rankgene: a program to rank genes from expression data," *Bioinformatics*, vol. 19, no.12, pp.1578-1579, Jan, 2003.

[6] A. Asuncion, and D. J. Newman, "UCI Machine Learning Repository [http://www.ics.uci.edu/mlearn/MLRepository .html]," irvine, CA: University of California, School of Information and Computer Science, 2007.



이 재 성

2007년 중앙대학교 컴퓨터공학과 졸업(학사). 2009년 중앙대학교 컴퓨터공학과 졸업(석사). 2009년~현재 중앙대학교 컴퓨터공학부 박사과정. 관심분야는 데이터마이닝, 패턴 인식, 패턴 분류, 특징 선별

김 대 원

정보과학회논문지 : 소프트웨어 및 응용
제 36 권 제 2 호 참조