한국통계학회 논문집
2010, 17권, 1호, 17-27

유전알고리즘을 이용한 최적 k-최근접이웃 분류기

박종선[1,a], 허균[a]

[a]성균관대학교 통계학과

요 약

분류분석에 사용되는 k-최근접이웃 분류기에 유전알고리즘을 적용하여 의미 있는 변수들과 이들에 대한 가중치 그리고 적절한 k를 동시에 선택하는 알고리즘을 제시하였다. 다양한 실제 자료에 대하여 기존의 여러 방법들과 교차타당성 방법을 통하여 비교한 결과 효과적인 것으로 나타났다.

주요용어: k-최근접이웃 분류기, 유전알고리즘, 변수선택, 변수가중치.

1. 서론

k-최근접이웃(k-Nearest Neighbor; k-NN) 분류기(classifier)는 초기에 형상인식분야에서 사용되기 시작하여 데이터마이닝 등의 분류문제에 폭넓게 활용되고 있다. 전형적인 k-NN 분류기는 범주형인 반응변수의 값을 모르는 관측치와 가장 근접한 k개의 표본을 훈련자료에서 추출하여 이들 중 가장 빈도가 높은 반응변수의 범주로 해당 관측치를 예측하는 비모수적인 방법이다. 이 방법은 그 동안 많은 연구가 이루어져 왔으며 다양한 실제 자료에서도 훌륭한 분류 성능을 갖는 것으로 알려져 있다 (Michie 등, 1994).

k-NN 분류기에서는 적절한 설명변수(feature)들의 선택과 이들에 대한 가중치가 분류기의 성능을 결정하는 중요한 요소들임이 여러 연구들 (Raymer 등, 2000; Paredes와 Vidal, 2000; Wettschereck 등, 1997; Bao 등, 2002)에서 밝혀졌으며 유전알고리즘을 이용하여 분류기에서 변수선택 및 가중치를 부여하는 다양한 방법들이 소개 되었다.

유전알고리즘을 이용한 분류기의 변수선택에 관한 연구는 1990년 Siedlecki와 Sklansky (1990)로부터 시작하여 k-NN 분류기를 사용하여 인쇄된 알파벳을 인식하는데 선택된 변수의 수를 24개에서 10개로 줄인 Smith 등 (1994)의 연구가 있으며 Fung 등 (1996)은 이를 서명인식에 적용하였다. Moser와 Murty (2000)는 직접 손으로 쓴 숫자의 인식을 위한 분류기에 적용하는 경우 효과적임을 보였다.

앞의 연구들에서 제시한 방법들을 적용하여 분류기에 필요한 적절한 변수들이 선택되었다 하더라도 각 변수들의 중요도가 모두 같을 수는 없다. 따라서 각 변수들에 최적의 가중치를 주는 방법에 대해서도 다양한 연구가 이루어졌다. Kelly와 Davis (1991)는 이진코딩이 아닌 실수코딩을 사용하여 분류정확도를 극대화하는 방법을 제시하였으며 Punch 등 (1993)은 이 연구를 확장하여 가중치의 범위를 제한하는 방법을 고려하였다. Komosiński와 Krawiec (2000)은 변수의 가중이 0 또는 1인 경우가 변수선택의 경우와 같은 의미를 갖는 점에 착안하여 이를 접목하였으며 Raymer 등 (2000)도 표준화한 각 변수들을 유전알고리즘을 이용하여 선택한 후 이들에 대한 가중치를 순차적으로 유전알고리즘을 이용하여 구하는 방법을 제시하였다. 또한 변수선택과 가중치의 부여를 동시에 고려하는 방법들에 대한 연구로 Tahir 등 (2007)이 혼성 타부(hybrid tabu) 탐색을 이용하는 방법을 제시하였다.

[1] 교신저자: (110-745) 서울시 종로구 명륜동3가 53번지, 성균관대학교 통계학과, 교수. E-mail: cspark@skku.edu

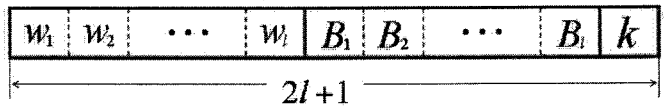

그림 1: *제안 알고리즘의 염색체 구성*

본 논문에서는 기존의 유전알고리즘을 이용한 k-NN 분류기의 최적화 방법에서 염색체의 표현 방식을 변형하여 k-NN의 분류성능 및 효율성을 높이기 위한 변수선택, 변수가중치 및 최근접 이웃 수 k의 선택을 동시에 고려하는 새로운 알고리즘을 제안하였다. 사용한 염색체 표현 방식은 Tahir 등 (2007)에 의해서 소개된 타부 탐색에서 사용된 인코딩 방법을 적용하여 염색체를 세 부분으로 구분하고 각각 다른 코딩방법을 사용하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 제안 알고리즘에 대한 상세한 설명을 포함하였으며 제 3장에서는 4개의 실제 자료에 대한 적용결과를 기존의 방법들과 비교하였다.

2. 제안 알고리즘

k-NN의 효율성을 높이기 위해 변수선택, 변수가중치 및 최근접 이웃수 k를 동시에 고려하는 염색체의 표현 방식은 기본적으로 변수선택과 변수가중치를 동시에 고려하는 Tahir 등 (2007)이 제시한 타부탐색 방법에서 사용한 인코딩 방법을 적용하였다.

우선 초기 l개의 설명변수들 중 x_s를 선택된 설명변수들의 집합이라 하고, x_{ij}는 i번째 패턴(관측치)의 j번째 설명변수의 값이라고 하자. 또한, 패턴(관측치)의 수는 n이고 유전알고리즘에서 모집단의 크기는 N으로 두었다.

2.1. 염색체(chromosome)의 구성

사용한 염색체는 그림 1과 같이 세부분으로 구성되었으며 염색체의 총길이는 초기변수 개수의 두 배에 1을 더한 값($2l+1$)으로 하였으며 각 부분에 대한 내용은 다음과 같다.

첫 번째, 변수선택을 위한 염색체(B_1에서 B_l까지)에서는 총 l개의 0 또는 1로 이뤄진 이진코딩(binary encoding)방법을 사용하였다. 이진코딩 염색체의 각 비트 1은 대응되는 변수가 선택됨을, 0은 선택되지 않음을 의미한다. 두 번째, 변수 가중을 위한 염색체(w_1부터 w_l까지)는 1.0부터 10.0까지의 실수 값의 열을 갖는 실수코딩(real encoding)방법을 사용했다. 총 l개의 실수값은 각각의 대응되는 변수의 가중치를 나타내며 변수선택 염색체의 값이 1인 변수들의 가중치만 사용된다. 세 번째, 최적의 k값을 찾기 위한 염색체 부분에는 1, 3, 5, 7 중 하나의 값을 갖도록 하였다. 이론적으로는 k값으로 모든 홀수가 가능하지만 이를 제한하여 7보다 작거나 같은 값으로 한정하였으나 필요에 따라 더 큰 값을 갖도록 할 수 있다.

2.2. 목적함수(objective function)

선택된 설명변수의 집합 x_s에 대한 목적함수 $f(x_s)$는 정확성과 효율성을 동시에 고려한 것으로 분류정확도와 선택된 변수의 개수를 동시에 고려하도록 하였다. 두 기준 사이의 조정계수는 0.01로 고정시켰으나 필요에 따라 다른 값을 사용할 수 있다.

$$f(x_s) = c(x_s) - \alpha \cdot \frac{\#(x_s)}{l} \tag{2.1}$$

x_s : 선택된 변수의 집합

W_1	W_2	.....	W_l	B_1	B_2	.....	B_l	k
1.5	2.7	.....	9.1	1	0	.....	0	1
4.2	3.1	.....	3.0	1	0	.....	0	3
8.1	2.8	.....	4.2	0	1	.....	1	5
⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋮
3.4	1.7	.....	5.7	0	0	.....	1	3
7.6	2.1	.....	6.1	1	1	.....	0	7

N

$2l+1$

그림 2: *초기 모집단의 형태*

$c(x_s)$: 검증자료에 대한 정분류율

α : 정분류율과 선택된 변수수 사이의 조정 계수($\alpha = 0.01$ 사용)

$\#(x_s)$: 선택된 변수의 개수

l : 설명변수 전체 개수

2.3. 초기모집단(initial population)

초기모집단을 생성하기 전에 편의(bias)가 발생하는 것을 방지하기 위해 아래의 식과 같이 모든 변수들을 [1.0, 10.0]의 범위를 갖는 값으로 정규화(normalize) 하였다.

$$x'_{i,j} = \left(\frac{x_{i,j} - \min\limits_{k=1,\ldots,n} \left(x_{k,j}\right)}{\max\limits_{k=1,\ldots,n} \left(x_{k,j}\right) - \min\limits_{k=1,\ldots,n} \left(x_{k,j}\right)} \times 9 \right) + 1 \tag{2.2}$$

$x_{i,j}$: i번째 관측치의 j번째 설명변수

$x'_{i,j}$: i번째 관측치의 j번째 설명변수의 정규화 값

n : 총 관측치의 수

난수발생기를 이용하여 초기 변수개수(l) 만큼의 이진정수를 생성하고, 마찬가지로 1.0부터 10.0까지의 가중치 범위에서 l개의 난수를 발생시켜 초기 가중치를 생성하였다. k의 초기값 역시 위와 같이 무작위로 설정하였다. 이 같은 작업을 모집단수(N) 만큼 반복하여 총 $(2l+1) \cdot N$ (총 염색체 길이 × 모집단 크기)개의 염색체로 구성된 초기모집단 $P(0)$을 생성한다. 생성된 초기모집단의 형태는 그림 2와 같다.

2.4. 재생산(reproduction)

재생산은 적자생존(survival of the fittest) 또는 자연도태(natural selection) 현상을 모방하려는 인위적인 메커니즘으로 이를 알고리즘으로 구현한 재생산 연산자는 적합함수값을 기반으로 $t(t = 0, 1, \ldots)$시점의 모집단 $P(t)$ 내의 개체들을 선택하고 교배급원을 형성해준다. 본 연구에서는 룰렛휠 선택방법을 이용하여 초기모집단으로부터 선택된 개체를 얻었다.

단계1) 먼저 t시점에서의 각 개체들의 적합도를 구하고 이를 모두 더한 적합도의 합을 계산한다.

$$f_{sum}(t) = \sum_{i=1}^{N} f_i(t), \tag{2.3}$$

$f_i(t)$는 t시점에서 i번째 개체의 적합도이고 N은 모집단크기.

단계2) 각 개체의 선택확률(selection probability) P_s를 계산한다.

$$P_s(s_i(t)) = \frac{f_i(t)}{f_{sum}(t)}, \quad (1 \le i \le N), \tag{2.4}$$

$s_i(t)$는 시점 t에서 i번째의 염색체.

단계3) 선택확률에 따라 개체를 선택하고 교배급원에 복제한다.

위의 단계3을 교배급원에 복제된 개체 수가 집단크기와 일치될 때까지 반복한다.

2.5. 교배(crossover)

교배는 탐색공간상의 가능한 새로운 점을 찾기 위하여 재생산 단계에서 얻은 교배급원으로부터 부모(parent) 염색체 쌍을 임의로 선택하고, 교배점 전후의 비트들을 서로 교환 및 결합함으로써 자손(offspring)을 생성한다. 변수선택과 변수가중에 관여하는 두 부분은 다음과 같은 일점교배 방법을 이용하여 교배하였다.

단계1) 교배급원으로부터 부모 염색체 쌍을 임의로 선정한다.

단계2) 교배확률(crossover rate) P_c를 토대로 교배유무를 결정한다. 난수 $r \in [0, 1]$를 발생시켜 만일 $r \le P_c$이면 교배를 일으킨다. 임의로 교배점 $c \in [1, l-1]$을 발생시키고, $[c+1, l]$사이의 유전자를 서로 교환함으로써 두 자손을 생성한다. 반대로 $r > P_c$이면 선택된 부모 염색체 쌍이 그대로 자손이 된다.

단계3) 생성된 자손을 $t+1$시점의 임시집단 $\tilde{P}(t+1)$에 복제한다.

임시집단이 N개의 염색체로 채워질 때까지 위의 연산은 반복된다. 근접 이웃수 k를 선택하기 위한 염색체의 세 번째 부분의 교배는 임의로 선택된 부모 염색체 쌍에서 k값이 다르다면 임의로 한 값을 선택하고 k값이 동일하다면 부모의 유전자가 그대로 자손으로 내려오는 교배방법을 이용하였다.

2.6. 돌연변이(mutation)

모의진화가 계속되는 동안 재생산과 교배 연산자는 집단을 더욱 동질적으로 만들고 이로 인하여 염색체들은 서로 닮아가게 된다. 이러한 현상이 세대 초기에 발생하면 유전자의 다양성 결핍으로 준 최적해(suboptimal solution)나 사점(dead corner)에 빠지게 되는 요인이 된다. 이를 방지하기 위하여 초기 세대에서 모든 염색체의 특정 비트가 고정되는 것을 방지해주고 또한 탐색영역을 확대해주는 돌연변이 연산자를 사용할 수 있다. 여기서는 가장 널리 이용되고 있는 방법인 단순(또는 표준) 돌연변이를 사용했으며 다음과 같이 3단계로 구성되어 있다.

단계1) 순차적으로 집단 $\tilde{P}(t+1)$내의 염색체에서 비트 하나를 취한다.

표 1: 실험 자료 요약 (단위: 개)

자료명	표본수	설명변수	범주수
Diabetes (Statlog)	768	8	2
Australian (Statlog)	690	14	2
Ionosphere (UCI)	351	34	2
Sonar (UCI)	208	60	2

Statlog: http://www.liaad.up.pt/ML/statlog/datasets.html
UCI: http://archive.ics.uci.edu/ml/index.html

단계2) 돌연변이 확률(mutation rate) P_m에 근거하여 선택된 비트의 돌연변이 유무를 결정한다. 난수 $r \in [0,1]$를 발생시켜 만약 $r \leq P_m$이면 돌연변이를 일으킨다. 이때 선택된 비트가 '1'이면 '0'으로, '0'이면 '1'로 반전시킨다. 반대로 $r > P_m$이면 비트 반전이 일어나지 않는다(염색체의 유전자들이 실수일 경우 돌연변이되면 지정한 영역 내에서 임의로 발생된 실수로 교체한다).

단계3) 선택된 비트를 집단 $P(t+1)$에 복제한다. 이 연산은 모든 염색체 인자들이 모두 체크될 때까지 반복되는데, 반복되는 루프는 $(2l+1) \cdot N$이 된다. 세대마다 확률적으로 돌연변이가 일어나는 비트 수는 약 $P_m \cdot (2l+1) \cdot N$개가 된다.

이와 같이 일련의 연산을 마치면 새로운 모집단 $P(t+1)$가 생성된다. 생성되는 모집단은 일반적으로 미리 정해진 최대 세대수 T에 도달할 때까지 반복된다.

3. 자료를 이용한 비교

이 장에서는 몇 가지 실제 자료를 이용하여 기존의 연구들에서 사용했던 방법들과 본 논문에서 제안한 유전알고리즘을 이용하여 변수선택 및 가중치, 그리고 k값을 동시에 고려하는 방법을 적용한 결과를 교차타당성 방법을 통하여 비교하였다. 모든 프로그램은 R을 이용하여 작성하였다.

3.1. 사용된 자료

분류 성능 비교를 위하여 Statlog와 UCI의 자료 중에서 설명변수의 수가 8, 14, 34, 60개인 4개의 자료(표 1)를 사용하였다.

각각의 자료에 대하여 살펴보면 Diabetes는 존스홉킨스 대학(The Johns Hopkins University)에서 제공한 피마족 인디언(Pima Indian)의 당뇨병 진단 자료로 임신횟수, 최소혈압, 혈청 인슐린, 나이 등을 포함한 8개의 변수들로 구성되어 있다. 모든 변수들은 연속형 변수이고 결측값(missing value)은 없다. Australian은 호주의 개인 신용카드와 관련된 자료이다. 모든 변수들의 이름과 값들은 기밀유지를 위해 의미 없는 값으로 변환되어 있으며 14개의 변수 중 범주형 변수 8개, 연속형 변수 6개를 포함하고 있다. Ionosphere 자료는 전리층으로부터 레이다의 신호가 돌아오는지 여부를 측정한 것이며 34개의 연속형 설명변수를 포함하고 있다. 마지막으로 Sonar는 다양한 각도와 조건에서 수중 음파탐지기를 하여 금속/비금속 실린더를 구별하는지 관찰한 자료이다. 자료의 수치들은 60개의 특정 주파수내에서의 에너지를 나타내며 0.0부터 1.0의 값을 가진다.

3.2. 실험 결과

제안 알고리즘을 적용하기 전에 모든 독립변수들을 식 (2.2)에 의해 표준화 하였으며 유전 알고리즘의 연산에 필요한 모수들은 교배확률을 80%, 돌연변이확률은 0.5%, 초기모집단수는 300, 반복세대

표 2: 제안 알고리즘의 분류 성능

자료명	반복	k	변수개수	목적함수값	오분류율(%)
Diabetes (8)	1	7	5	0.80	19.9
	2	7	5	0.79	20.8
	3	7	6	0.80	19.5
	4	7	5	0.79	20.3
	5	7	8	0.80	19.0
				Average	**19.9**
Australian (14)	1	5	6	0.89	10.1
	2	5	6	0.89	10.6
	3	5	8	0.89	9.6
	4	5	8	0.89	9.6
	5	5	6	0.89	10.1
				Average	**10.0**
Ionosphere (34)	1	1	8	0.97	2.8
	2	3	6	0.97	2.8
	3	1	11	0.97	2.8
	4	1	14	0.97	2.8
	5	1	14	0.96	3.8
				Average	**3.0**
Sonar (60)	1	1	32	0.98	1.6
	2	1	29	0.98	1.6
	3	1	24	0.98	1.6
	4	1	22	0.98	1.6
	5	1	26	0.98	1.6
				Average	**1.6**

수는 500으로 주었다. 단, Sonar 자료의 경우 과잉적합을 피하기 위해 반복세대수를 300으로 설정하였다. 이 값들은 기존의 연구결과와 비교를 위하여 Tahir 등 (2007)과 Kudo와 Sklansky (2000)에서 사용된 값을 그대로 적용하였다. 반복세대수의 경우에는 자료별로 반복 실험을 통하여 수렴여부를 판단한 후 수렴하도록 조정하였다.

각각의 자료들은 70%로 구성된 훈련자료(train data)와 나머지 30%로 구성된 검증자료(test data)로 임의로 분할하였다. 각 자료마다 5번씩의 반복을 통하여 최적으로 추정된 k, 선택된 변수개수, 목적함수 값, 검증자료의 오분류율(%)을 살펴보았으며 그 결과는 표 2와 같다. 모든 자료에서 적합도 값이 비교적 일정하게 수렴하는 것을 볼 수 있으며 최적의 k값도 하나의 값으로 수렴하였다. 하지만 선택된 변수의 개수와 변수별 가중치에는 차이가 있는 것으로 나타나 다중 해가 존재함을 알 수 있다.

표 3은 각각의 자료에 대해 제안 알고리즘과 다른 분류기들의 오분류율(%)을 비교한 것이다. 비교에 사용된 분류기들은 결정나무 방법인 C4.5 (Quinlan, 1993)와 일반적인 k-NN 분류기와 Tahir (2007)의 타부탐색 k-NN 분류기(TS/k-NN), 선형 판별분석, 베이지안 방법인 나이브 베이즈분류기(Naive Bayes classifier)이며 제안된 방법 이외의 모든 방법들에서는 설명변수들을 모두 사용하였다.

실험 결과를 보면 k-NN에서 유전 알고리즘을 이용하여 최적화 시킨 제안 알고리즘의 분류정확도가 모든 자료에서 우수한 성능을 보였다. 사용한 오분류율은 각 자료마다 5회 반복 실험한 결과의 평균 오분류율이다. 변수의 개수가 적은 Diabetes 자료의 경우 모든 분류기의 성능이 비슷하지만, 변수의 개수가 각각 34, 60개인 Ionosphere, Sonar 자료에서는 유전 알고리즘과 타부탐색법을 k-NN에 결합한 두 분류기(제안 방법, TS/k-NN)의 성능이 월등히 우수했다. 특히 Ionosphere 자료에서 제안 알고리즘의 오분류율은 3.0%, Sonar 자료에서 제안 알고리즘의 오분류율은 1.6%로써 타부탐색법을 이용한 분

표 3: 제안 알고리즘의 오분류율 비교 (단위: %)

자료명	C4.5	k-NN	판별분석	베이즈	TS/k-NN	제안방법
Diabetes	26.3	29.7	22.5	23.5	22.3	**19.9**
Australian	15.6	16.7	14.1	20.8	10.2	**10.0**
Ionosphere	11.3	14.3	13.1	19.1	6.2	**3.0**
Sonar	23.1	12.5	25.0	22.5	5.8	**1.6**

표 4: 제안 알고리즘의 분류 성능 비교

자료명	방법	오분류율(%)	변수개수	k
Diabetes (8)	SFS	23.7	4	7
	SFFS	23.7	4	7
	TS(FS)	23.7	4	7
	TS(FW)	22.3	8	7
	TS(FS+FW)	20.1	4	5
	제안방법	**19.0**	**8**	**7**
Australian (14)	SFS	11.6	6	7
	SFFS	11.3	5	9
	TS(FS)	11.3	5	9
	TS(FW)	9.1	14	7
	TS(FS+FW)	7.1	9	3
	제안방법	**9.6**	**8**	**5**
Ionosphere (34)	SFS	8.5	11	3
	SFFS	7.5	9	3
	TS(FS)	6.6	9	3
	TS(FW)	8.5	34	1
	TS(FS+FW)	5.7	15	3
	제안방법	**2.8**	**6**	**3**
Sonar (60)	SFS	7.2	35	1
	SFFS	4.8	33	1
	TS(FS)	3.4	24	1
	TS(FW)	6.7	60	1
	TS(FS+FW)	5.8	17	1
	제안방법	**1.6**	**22**	**1**

류보다도 우수한 결과를 보였다. 위 표에서 C4.5, k-NN, 판별분석, 나이브베이즈 방법을 이용한 분류는 각 자료의 모든 변수를 이용한 분류결과이므로 최소의 변수 개수로 최소의 오분류율을 보인 제안 알고리즘의 효율성 또한 매우 우수함을 알 수 있다.

표 4는 제안 알고리즘과 k-NN에 순차전진탐색(SFS; Sequential Forward Search) 방법과 SFS 방법을 개선한 순차전진 플로팅탐색(SFFS; Sequential Forward Floating Search) 그리고 변수선택(FS; Feature Selection)과 변수가중치(FW; Feature Weighting)를 적용한 타부탐색 k-NN을 비교한 결과를 포함하고 있다. 표에서의 오분류율(%)과 선택된 변수개수, k값은 5번의 실험 중 적합도가 가장 높은 시행에서의 수치들이다.

먼저, 제안 알고리즘을 이용한 분류기의 분류성능은 Australian 자료를 제외한 세 자료에서 가장 우수함을 알 수 있으며 특히, 변수의 개수가 많은 자료(Ionosphere, Sonar)에서 다른 변수선택방법보다 상대적으로 매우 작았다. 선택된 변수의 개수도 다른 방법과 비슷하거나 작으므로 효율성도 우수한 것으로 볼 수 있다. 추정된 최적의 k값은 Australian 자료를 제외하면 다른 방법들과 동일하였다. 변수선택방법들 간의 직접적인 비교를 위하여 표 4의 오분류율과 변수개수를 이용하여 목적함수값을 계산하였고 그 결과는 그림 3과 같다. Australian 자료를 제외한 세 개의 자료에서 다른 변수선택방법들에 비해

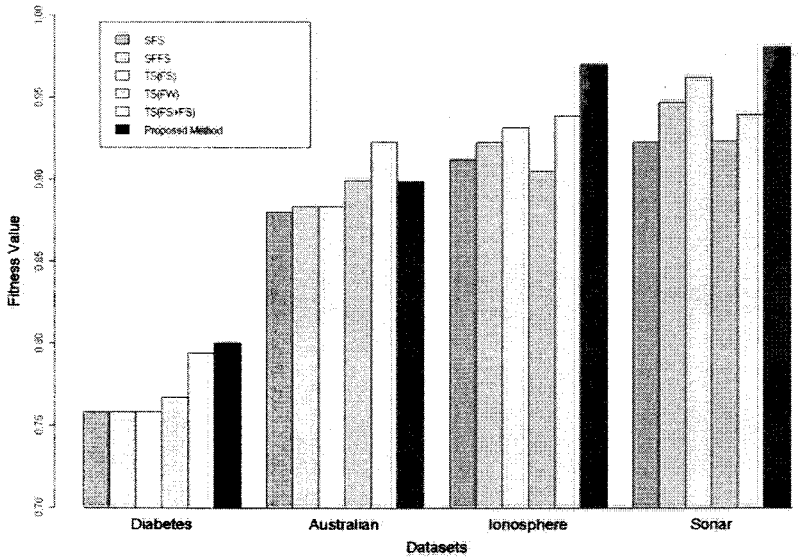

그림 3: *변수선택 방법들 간의 적합도 비교*

제안 알고리즘을 이용한 분류방법의 목적함수값이 높은 것으로 나타나 목적함수의 측면에서도 효율적임을 알 수 있다.

변수의 개수에 따라 자료를 소규모[0, 19], 중규모[20, 49], 대규모[50, ∞]로 구분했을 때 Diabetes와 Australian 자료는 소규모, Ionosphere 자료는 중규모, Sonar는 대규모 자료에 해당한다. 서론에서 언급한 것과 같이 SFS, SFFS와 같은 순차적 알고리즘은 소규모 자료(Diabetes, Australian)에서 다른 분류방법과 비슷한 성능을 보였지만, 변수개수가 큰 중/대규모 자료(Ionosphere, Sonar)에서는 지역 최적해에 빠지는 문제가 발생하므로 타부탐색과 유전 알고리즘과 같은 휴리스틱 알고리즘을 이용한 분류기의 성능이 우수했다. Ionosphere 자료에서는 제안 알고리즘의 경우 34개의 변수 중에 단 6개의 선택된 변수로써 97.2%의 높은 분류정확도를 보였고 Sonar 자료에서는 60개의 변수 중에 22개의 선택된 변수로써 98.4%의 높은 분류 결과를 보였다.

4. 결론

본 연구에서는 우수한 전역탐색기능과 뛰어난 정확도를 가지는 유전 알고리즘을 이용하여 k-NN에서 최적의 변수선택 및 가중치, k값의 선택을 동시에 하는 분류방법을 제안하였다. 전체적으로 다양한 실제 자료에 대하여 적용한 결과 자료의 규모에 상관없이 제안 알고리즘을 이용한 분류방법이 기존의 여러 분류기들에 비하여 분류정확도 및 효율성이 우수한 것으로 나타났다. 제안 알고리즘에서 고려한 변수선택 및 가중치 그리고 k 중에서 k값은 Austrailian 자료를 제외하면 대부분의 경우에 기존의 방법들과 같은 값이 선택되어 상대적으로 중요성이 낮은 것으로 보인다. 하지만 같은 k값의 경우에도 제안 알고리즘이 기존의 방법들에 비하여 효율적인 것으로 나타나 선택된 변수들과 이들에 대한 가중치의 차이가 분류기의 성능에 많은 영향을 주는 것을 알 수 있다.

기존의 방법들과의 비교를 위하여 목적함수에서 오분류율과 변수개수 사이의 조정계수 α값을 0.01로 고정한 결과만을 포함하였으나 계수 값에 따라 최적의 변수, 가중치 그리고 k값에 많은 차이가 있을 것으로 예상된다. 조정계수 α값 및 여타 유전알고리즘의 모수들에 대한 다양한 시뮬레이션을 통

하여 주어진 자료에 최적인 모수들의 조율이 가능할 것이다. 하지만 제안 알고리즘은 여러 가지를 동시에 고려하는 방법인 만큼 수렴속도도 늦고 각 자료마다 지역최적해에 빠지지 않고 전역최적해에 수렴하기 위한 반복세대수, 모집단 크기, 재생산을 위한 방법 선택, 교배방법의 선택, 돌연변이 비율 등을 결정하는 것은 간단하지 않을 것으로 보인다.

끝으로, 유전 알고리즘은 전역적 탐색방법으로 전체 공간에 대한 탐색을 통해 효율적으로 최적해에 수렴하는 특성을 갖고 있고, 타부 탐색법은 국소적 탐색방법으로 최적해 근방에서 시작하면 효율적으로 최적해에 수렴하는 특성을 갖고 있으므로 이 두 가지 방법을 적절히 결합시키는 방법도 연구할 가치가 있을 것이다.

참고 문헌

Bao, Y., Du, X. and Ishii, N. (2002). *Combining Feature Selection with Feature Weighting for k-NN Classifier*, IDEAL 2002, LNCS 2412. Springer-Verlag, 461–468.

Fung, G., Liu, J. and Lau, R. (1996). Feature selection in automatic signature verification based on genetic algorithms, In *Proceedings of International Conference on Neural Information*, 811–815.

Kelly, J. and Davis, L. (1991). A hybrid genetic algorithm for classification, In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 645–650.

Komosiński, M. and Krawiec, K. (2000). Evolutionary weighting of image features for diagnosing of CNS tumors, *Artificial Intelligence in Medicine*, **19**, 25–38.

Kudo, M. and Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifier, *Pattern Recognition*, **33**, 25–41.

Michie, D., Spiegelhalter, D. and Taylor, C. (1994). *Machine Learning Neural and Statistical Classification*, Ellis Horwood.

Moser, A. and Murty, M. (2000). On the scalability of genetic algorithms to very large-scale feature selection, *Real World Applications of Evolutionary Computing*, 77–86.

Paredes, R. and Vidal, E. (2000). A Class-Dependent Weighted Dissimilarity Measure for Nearest Neighbor Classification Problems, *Pattern Recognition Letert*, **21**, 1027–1036.

Punch, W., Goodman, E. and Pei, M. (1993). Further research on feature selection and classification using genetic algorithms, In *Proceedings of the Fifth International Conference on Genetic Algorithms*, 379–383.

Quinlan, J. (1993). C4.5: Programs for machine learning, Morgan Kaufmann.

Raymer, M., Punch, W., Goodman, E., Kuhn, L. and Jain, A. (2000). Dimensionality reduction using genetic algorithms, *IEEE Transactions on Computers*, **4**, 164–171.

Siedlecki, W. and Sklansky, J. (1990). A note on genetic algorithms large-scale feature selection, *IEEE Transactions on Computers*, 335–347.

Smith, J., Fogarty, T. and Johnson, I. (1994). Genetic feature selection for clustering and classification, In *Proceedings IEE Colloquium Genetic Algorithms in Image Processing Vision*, 193–196.

Tahir, M., Bouridane, A. and Kurugullu, F. (2007). Simultaneous feature selection and feature weighting using Hybrid Tabu Search/k-nearest neighbor classifier, *Patter Recognition Letter*, **28**, 438–446.

Wettschereck, D., Aha, D. W. and Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms, *Artificial Intelligence Review*, **11**, 273–314.

2009년 11월 접수; 2009년 12월 채택

Optimal k-Nearest Neighborhood Classifier Using Genetic Algorithm

Chongsun Park[1,a], Kyun Huh[a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

Feature selection and feature weighting are useful techniques for improving the classification accuracy of k-Nearest Neighbor (k-NN) classifier. The main propose of feature selection and feature weighting is to reduce the number of features, by eliminating irrelevant and redundant features, while simultaneously maintaining or enhancing classification accuracy. In this paper, a novel hybrid approach is proposed for simultaneous feature selection, feature weighting and choice of k in k-NN classifier based on Genetic Algorithm. The results have indicated that the proposed algorithm is quite comparable with and superior to existing classifiers with or without feature selection and feature weighting capability.

Keywords: k-Nearest Neighborhood classifier, genetic algorithm, feature selection, feature weighting.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 53 Myungnyun-Dong, Jongno-Gu, Seoul 110-745, Korea. E-mail: cspark@skku.edu