인식기 풀 기반의 다수 인식기 시스템 구축 방법

(Construction of Multiple Classifier Systems based on a Classifiers Pool)

강 희 중 †
(Hee-Joong Kang)

요 약 우수한 인식 성능을 보이기 위하여 가용한 인식기 풀(pool)로부터 다수 인식기를 선택하는 방법에 관한 연구는 소수에 불과하였다. 그래서, 어떻게 또는 얼마나 많은 인식기를 선택해야 하는가에 관한 인식기의 선택 문제는 여전히 중요한 연구 주제로 남아 있다. 본 논문에서는 선택되는 인식기의 개수가 미리 제한되어 있다는 가정 하에서, 다양한 선택 기준을 제안하고, 이들 선택 기준에 따라서 다수 인식기 시스템을 구축하며, 구축된 다수 인식기 시스템의 성능을 평가함으로써 제안된 선택 기준을 평가하고자 한다. 모든 가능한 다수 인식기의 집합은 선택 기준에 의해서 조사되고, 그 중 일부가 다수 인식기 시스템의 후보로 선정된다. 이러한 다수 인식기 시스템 후보들은 Concordia 대학과 UCI(University of California, Irvine)의 기계학습 자료로부터 얻은 무제약 필기 숫자를 인식하는 실험에 의해 평가되었다. 다양한 선택 기준 중에서, 특히 조건부 엔트로피에 기반한 정보 이론적 선택 기준에 의하여 구축된 다수 인식기 시스템 후보가 다른 선택 기준에 의한 후보보다 더 유망한 결과를 보여 주었다.

키워드 : 인식기 풀, 다수 인식기 시스템, 정보 이론, 유사도, 조건부 엔트로피, 의존관계, 베이지안 방법

Abstract Only a few studies have been conducted on how to select multiple classifiers from the pool of available classifiers for showing the good classification performance. Thus, the selection problem of classifiers on how to select or how many to select still remains an important research issue. In this paper, provided that the number of selected classifiers is constrained in advance, a variety of selection criteria are proposed and applied to the construction of multiple classifier systems, and then these selection criteria will be evaluated by the performance of the constructed multiple classifier systems. All the possible sets of classifiers are examined by the selection criteria, and some of these sets are selected as the candidates of multiple classifier systems. The multiple classifier system candidates were evaluated by the experiments recognizing unconstrained handwritten numerals obtained both from Concordia university and UCI machine learning repository. Among the selection criteria, particularly the multiple classifier system candidates by the information-theoretic selection criteria based on conditional entropy showed more promising results than those by the other selection criteria.

Key words : a pool of classifiers, multiple classifier system, information theory, measare of closeness, conditional entropy, dependency, Bayesian method

1. 서 론

인식 성능을 향상시키기 위하여 다수 인식기를 결합하려는 연구가 지난 10여 년간 문자 인식의 분야에서 활발히 수행되어 왔다. 이들 연구는 주로 개별 인식기의 인식 결과를 결합하는 것에 관한 결합 방법에 중점을 두었으며, 제안된 여러 결합 방법들은 개별 인식기의 성능에 비해서 상당히 향상되었음을 보고해 왔다[1,2,3,4]. 그렇지만, 가용한 인식기 풀(pool)로부터 어떻게 다수 인식기 시스템을 구축하는가에 관한 연구는 소수에 불과했다[5,6]. 그러므로, 어떻게 또는 얼마나 많은 인식기

· 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2000-003-E00296).
† 종신회원 : 한성대학교 컴퓨터공학부 교수
hjkang@hansung.ac.kr
논문접수 : 2001년 11월 2일
심사완료 : 2002년 6월 19일

를 선택해야 하는가에 관한 인식기의 선택 문제는 여전히 중요한 연구 주제로 남아 있다. 다수 인식기를 결합함으로써 우수한 인식 성능을 이루는 것은 결합 방법은 물론, 인식기 풀로부터의 인식기의 선택에도 의존한다.

이러한 인식기 선택 문제를 완화시켜 해결하기 위하여 다수 인식기 시스템에 사용될 인식기의 개수가 미리 제한되어 있다고 가정한다. 본 논문에서는, 인식기의 인식 성능에만 기반한 단순한 선택 기준에 덧붙여서, [7,8,9]에 있는 정보 이론에 근거한 정보 이론적 기준을 제안하고, 제안된 선택 기준에 의하여 다수 인식기 시스템을 구축하고자 한다. 단순한 선택 기준 중의 하나는 인식기의 제한된 개수까지 우수한 정인식율의 순서대로 인식기를 선택하는 것이고, 다른 하나의 기준은 인식기의 선택에 있어서 정인식율 대신에 신뢰율을 사용하는 것이다. 이들 단순한 선택 기준은 필기 숫자 인식 실험에서 정보 이론적 기준과 비교된다. 정보 이론적 기준의 하나는 [7]에서 정의된 유사도(Measure of Closeness) 기준을 사용하는 것이고, 다른 하나는 [9]에서 정의된 조건부 엔트로피 (Conditional Entropy)를 사용하는 것이다.

본 논문에서 제안되는 다양한 인식기 선택 기준은 인식기 풀에 적용되어 제한된 인식기의 개수에 따라 인식기의 가능한 집합을 구성하고, 개별 인식기 집합을 검사한다. 그 다음, 1등만을 고려할 경우, 다수 인식기 시스템의 후보로서 한 인식기 집합을 선정한다. 다수 인식기 시스템 후보는 *Concordia* 대학의 CENPARMI 연구소 [10]와 UCI (University of California, Irvine)의 기계학습 자료[11]에서 얻은 무제약 필기 숫자를 인식하는 실험에서 다른 인식기 집합과 함께 [1,2,4,12]에서 제안된 결합 방법에 의해서 평가되었다. 제안된 조건부 엔트로피에 기반한 정보 이론적 기준에 의해서 선정된 다수 인식기 시스템 후보가 실험에서 가장 우수한 다수 인식기 시스템과 항상 일치하지는 않는다 하더라도 다른 선택 기준에 의한 다수 인식기 시스템 후보 보다 더 우수한 인식 결과를 보여 주었고, 특히 5개 인식기를 결합하는 경우에 있어서는 인식기 선택에 관해 유망한 결과를 보여 주었다. 무제약 필기 숫자 인식 실험을 위해서, 단일 선택 또는 순위 레벨의 결정을 수행하는 6개의 인식기를 [13,14]와 같이 훈련시켜 인식기 풀에 두었다.

본 논문의 구성은 다음과 같다. 2장에서는 인식기의 선택과 관련된 기존 연구를 살펴보고, 3장은 인식기를 선택하기 위한 다양한 선택 기준을 소개하며, 인식기 선택 기준의 평가를 위하여 Concordia 대학과 UCI의 기계학습 자료로부터 얻은 무제약 필기 숫자에 대한 인식 실험을 4장에서 기술하고, 5장에서 결론을 맺는 순서로 되어 있다.

2. 관련 연구

다양한 인식기 집합이 주어졌을 때, 최적의 결합 결과를 얻기 위하여 어떤 인식기 집합을 결정해야 하는가에 관한 방법론을 위해서, 김종렬 등은 [5]에서 개별 인식기로부터 발생된 오류로부터 계산될 수 있다고 여긴 인식기간의 친밀도를 제안하였다. 그렇지만, 친밀도는 체계적으로 확장성 있게 계산되지 않았고, 친밀도 계산을 위한 관련 파라미터도 쉽게 결정할 수 없다는 단점을 지니고 있었다. Woods 등은 [3]에서 인식기 집합을 결정할 때 어떤 전략이 필요하다고 제안하였다. 왜냐하면, 모든 데이타 집합에 대해서 5개 인식기의 결합 결과 보다 더 우수한 결과가 4개 인식기의 결합에서 존재하기 때문이다. 이들은 인식기 집합을 선정하는데 있어서 간단한 순차적 후향 탐색을 고려하였고, 인식기의 지역적 정확율 측정치를 이용하여 어떤 인식기가 중복되거나 해로운 지를 판단하는 프로시저를 제안하였다. 이들의 접근 방법 또한 단순하고 체계적이지 못한 단점을 지니고 있다. Impedovo와 Salzo는 [6]에서 미리 지정된 결합 방법으로 결합할 인식기 집합을 선정하기 위한 다수 인식기 시스템의 설계에 관해서 연구하여 결합 방법의 평가를 위한 새로운 방법을 제안하였다. 이 방법은 인식기의 정인식율과 친밀지수로 측정되는 인식기간의 상관관계를 고려하여 다수 인식기 시스템을 설계하는 것이다. 이러한 기존의 관련 연구는 대체로 1차 의존관계만을 고려하여 인식기의 집합을 선정하는데 있어서 친밀도나 친밀지수에 활용하였다.

3. 인식기의 선택 기준

인식 성능의 향상을 위해서는 개별 인식기들이 서로 상호 보완적이도록 다수 인식기 시스템을 구축하는 것이 바람직하다. 이러한 목적으로 인식기의 선택 기준은 종종 개별 인식기의 성능은 물론, 다수 인식기 시스템에서 사용될 결합 방법도 고려하게 된다. 인식기가 독립적으로 수행한다는 가정은 인식기가 통계적으로 다른 것들에 의존적인 성향을 따르기 때문에 취약하다고 볼 수 있다. 따라서, 본 논문에서는 인식기가 독립이라는 가정하에서의 결합 방법들과 그 가정을 필요로 하지 않는 결합 방법들을 모두 고려하여 인식기의 선택 기준에 대한 평가를 수행하고자 한다.

본 논문에서는, 앞서 기술한대로 두 가지의 단순한 선택 기준을 제안하였다. 하나는 FRR(Forced Recognition Rate) 기준인데, 이는 인식기를 선정하는데 있어서 정인식율을 사용하는 것이고, 반면에 다른 하나는 RR

(Reliability Rate) 기준이며, 정인식율 대신에 신뢰율을 사용하는 것이다. 또한 이러한 단순한 선택 기준에 덧붙여서, 인식기간의 1차와 2차 의존관계를 고려하는 두 가지의 정보 이론적 기준을 제안하고자 한다. 이들 의존관계는 베이지안 결합 방법에서 고차 확률 분포를 저차 분포의 곱으로 근사하기 위한 이론적 근거를 제공하고 있다. 정보 이론적 기준의 하나는 [7]에서 정의된 유사도에 근거를 두고 있으므로 MC (Measure of Closeness) 기준이라 정하고, 다른 하나는 [9]에서 정의된 베이스 에러율의 상위 경계의 최소화에 기준을 두고 있어서 CE (Conditional Entropy) 기준이라 정한다.

변수 C가 클래스 변수 C_{K+1}와 K개 인식기의 인식 결과 변수 $C_1,\cdots,C_K$들을 표현하는 $(K+1)$차원 변수라고 할 때, 유사도는 실제 확률 분포인 $P(C)$와 이의 근사 확률 분포인 $P_a(C)$의 유사성에 대한 차이를 계산하는 기준으로, 두 확률 분포가 동일하면 차이가 없게 되어 0의 값을 지니게 된다. 따라서, 이러한 유사성에 대한 차이를 최소화함으로써 최적의 근사 확률 분포를 얻을 수 있으며, 두 확률 분포간의 유사성을 높이기 위한 기준으로서 적용된다. 이러한 유사도 $I(P(C),P_a(C))$은 변수 C가 가질 수 있는 값을 c이라 할 때 다음 식과 같이 정의된다[7].

$$I(P(C),P_a(C))=\sum_c P(c)\log\frac{P(c)}{P_a(c)} \tag{1}$$

변수 C의 $(K+1)$차 확률 분포에 대한 곱 근사 확률 분포를 얻기 위하여 d차 의존관계를 고려한 곱 근사 분포 식은 다음과 같이 정의된다.

$$P_a(C_1,\cdots,C_{K+1})=\prod_{j=1}^{K+1}P(C_{n_j}\mid C_{n_{id(j)}},\cdots,C_{n_{il(j)}}),(0\le id(j),\cdots,il(j)<j) \tag{2}$$

위 식(2)에서 C_{n_j}는 $C_{n_{id(j)}}$과 $C_{n_{il(j)}}$까지 d개의 항들에 조건부로 정의되고, $(n_1,\ldots,n_K,n_{K+1})$와 이의 d개 조건부 미지 순열인 $(n_{id(1)},\ldots,n_{id(K+1)})\cdots(n_{il(1)},\ldots,n_{il(K+1)})$ 등은 1부터 $K+1$까지의 정수에 대한 미지 순열을 의미한다. 정의에 의하여 $C_{n_{j-(j)}}$이 $C_{n_{id(j)}}$와 $C_{n_{il(j)}}$사이의 한 항을 지칭할 때 $P(C_{n_j}\mid C_0,C_{n_{j-(j)}})$은 $P(C_{n_j},C_{n_{j-(j)}})$이라 정하고, C_0은 널(null) 항을 의미한다. 식(2)의 우변과 같은 곱 근사 분포는 다양하게 존재할 수 있는데, 우리가 관심이 있는 것은 최적인 즉, 유사성에 대한 차이가 최소인 곱 근사 분포이다. 이러한 최적의 곱 근사 분포를 구하는 이론적 근거와 알고리즘은 [4]에 자세히 기술되어 있으며, 최적의 곱 근사 분포를 구하는 것은 식(1)과 식(2)로부터 유도된 구성 항들의 상호정보의 합이 최대인 곱 근사 분포를 결정하는 것과 같다. 요약하면, MC 기준은 가능한 각 인식기 집합에 대하여 식(1)과 식(2)을 기준으로 주어진 의존관계 차수로 표현할 수 있는 최적의 곱 근사 분포를 구하고, 이러한 최적의 곱 근사 분포들이 지닌 상호정보의 합들 중에서 최대 상호정보의 합을 지닌 곱 근사 분포와 관련된 인식기 집합을 선택하는 것이다. 이를 간략히 알고리즘으로 표현하면 다음과 같다.

입력 : 의존관계 차수 d와 인식기의 개수 제한을 만족하는 w개의 인식기 집합 $S^1,S^2,\ldots,S^w$

출력 : 최대 상호정보의 합을 지닌 인식기 집합 S^*

방법 : 1. 개별 인식기 집합에 대하여 샘플로부터 의존관계 차수에 따라 계산된 상호정보의 합이 최대인 곱 근사 분포와 그 합을 [4]의 알고리즘을 이용하여 구한다.

2. 모든 인식기 집합들에 대하여 1단계에서 계산된 상호정보의 합들로부터 최대 상호정보의 합을 구하고, 그 합에 관련된 인식기 집합을 MC 기준에 의해서 선택된 인식기 집합으로 결정하여 출력한다.

한편, 베이스 에러율의 상위 경계를 최소화하는 방법은 [12]에서와 같이 클래스 변수인 M과 K개 인식기의 인식 결과를 표현하는 K차원 변수인 E로 구성된 조건부 엔트로피 $H(M|E)$를 최소화하여 베이지안 결정에서의 에러율을 최소화하려는 것인데, [12]에서 정의되었고 여기에도 사용되는 클래스 변수와 인식 결과 변수에 관한 CD 상호정보 $U(M;E)$에 이들 변수의 확률 분포를 적용하여 CD 상호정보를 최대화시키면, 앞서 기술한 $(K+1)$차 확률 분포에 대한 최적의 근사 확률 분포를 얻을 수 있게 된다. 베이스 에러율 P_e의 정의는 변수 M과 E가 가질 수 있는 값을 각각 m과 e라고 할 때 아래 식(3)과 식(4)에서와 같은 식으로 정의된다[9].

$$P_e\le\frac{1}{2}H(M\mid E)=\frac{1}{2}(H(M)-U(M;E)) \tag{3}$$

$$U(M;E)=\sum_m\sum_e P(m,e)\log\frac{P(m,e)}{P(m)P(e)} \tag{4}$$

위와 같은 방법으로 변수 M과 E의 $(K+1)$차 확률 분포에 대한 곱 근사 확률 분포를 얻기 위하여 인식기만의 d차 의존관계를 고려한 곱 근사 분포 식은 다음과 같이 정의된다.

$$P_a(E_1,\cdots,E_K,M)=\prod_{j=1}^{K}P(E_{n_j}\mid E_{n_{id(j)}},\cdots,E_{n_{il(j)}},M),\quad (0\le id(j),\cdots,il(j)<j) \tag{5}$$

$$P_a(E_1,\cdots,E_K)=\prod_{j=1}^{K}P(E_{n_j}\mid E_{n_{id(j)}},\cdots,E_{n_{il(j)}}),\quad (0\le id(j),\cdots,il(j)<j) \tag{6}$$

위 식(5)에서 E_{n_j}는 $E_{n_{i2(j)}}$과 $E_{n_{i2(j)}}$까지 d개의 항들과 클래스 변수인 M에 조건부로 정의되고, 식(6)에서 E_{n_j}는 $E_{n_{i1(j)}}$과 $E_{n_{i1(j)}}$까지 d개의 항들에 조건부로 정의된다. 또한, $(n_1, \ldots, n_K)$와 이의 d개 조건부 미지 순열인 $(n_{i2(1)}, \ldots, n_{i2(K)})\cdots(n_{i1(1)}, \ldots, n_{i1(K)})$ 등은 1부터 K까지의 정수에 대한 미지 순열을 의미한다. 정의에 의하여 $E_{n_{i-1,j}}$이 $E_{n_{i2(j)}}$와 $E_{n_{i1(j)}}$사이의 한 항을 지칭할 때, 식(5)에서 $P(E_{n_j} \mid E_0, E_0, M)$은 $P(E_{n_j}, M)$이라 정하고, $P(E_{n_j} \mid E_0, E_{n_{i-1,j}}, M)$은 $P(E_{n_j} \mid E_{n_{i-1,j}}, M)$이라 정하며, 식(6)에서 $P(E_{n_j} \mid E_0, E_{n_{i-1,j}})$은 $P(E_{n_j} \mid E_{n_{i-1,j}})$이라 정하고, E_0은 공히 널(null) 항을 의미한다. 식(5) 또는 식(6)의 우변과 같은 곱 근사 분포는 다양하게 존재할 수 있는데, 우리가 관심이 있는 것은 최적인 즉, 식(4)와 같은 CD 상호정보가 최대인 곱 근사 분포를 식(5)와 식(6)과 같은 형태로 구하는 것이다. 이러한 최적의 곱 근사 분포를 구하는 이론적 근거와 알고리즘은 [12]에 자세히 기술되어 있으며, 최적의 곱 근사 분포를 구하는 것은 식(5)와 식(6)을 식(4)에 적용함으로써 유도된 구성 항들의 상호정보의 차이의 합이 최대인 곱 근사 분포를 결정하는 것과 같다. 상호정보의 차이는 [12]에서 정의된 바와 같이 Δ상호정보라고 정의한다. 참고로, 2차 의존관계에 의한 Δ상호정보는 식(7)과 같고, 인식기만을 고려한 2차 의존관계에 의한 상호정보는 클래스 변수의 유무에 따라서 각각 식(8) 또는 식(9)와 같다.

$$\Delta D(E_j; E_{i2(j)}, E_{i1(j)}) = D(E_j; E_{i2(j)}, E_{i1(j)}, M) - D(E_j; E_{i2(j)}, E_{i1(j)}) \quad (7)$$

$$D(E_j; E_{i2(j)}, E_{i1(j)}, M) = \sum_E \sum_M P(E, M) \log P(E_j \mid E_{i2(j)}, E_{i1(j)}, M) \quad (8)$$

$$D(E_j; E_{i2(j)}, E_{i1(j)}) = \sum_E P(E) \log P(E_j \mid E_{i2(j)}, E_{i1(j)}) \quad (9)$$

요약하면, CE 기준은 가능한 각 인식기 집합에 대하여 식(4)를 기준으로 식(5)와 식(6)을 적용하여 주어진 의존관계 차수로 표현할 수 있는 최적의 곱 근사 분포를 구하고, 이러한 최적의 곱 근사 분포들이 지닌 Δ상호정보의 합들 중에서 최대 Δ상호정보의 합을 지닌 곱 근사 분포와 관련된 인식기 집합을 선택하는 것이다. 이를 간략히 알고리즘으로 표현하면 다음과 같으며, 베이스 에러율의 상위 경계의 최소화에 따른 조건부 엔트로피 사용에 대한 자세한 내용은 [12]에 소개되어 있다.

입력 : 의존관계 차수 d와 인식기의 개수 제한을 만족하는 w개의 인식기 집합 $S^1, S^2, \ldots, S^w$

출력 : 최대 Δ상호정보의 합을 지닌 인식기 집합 S^*

방법 : 1. 개별 인식기 집합에 대하여 샘플로부터 의존관계 차수에 따라 계산된 Δ상호정보의 합이 최대인 곱 근사 분포와 그 합을 [12]의 알고리즘을 이용하여 구한다.

2. 모든 인식기 집합들에 대하여 1단계에서 계산된 Δ상호정보의 합들로부터 최대 Δ상호정보의 합을 구하고, 그 합에 관련된 인식기 집합을 CE 기준에 의해서 선택된 인식기 집합으로 결정하여 출력한다.

4. 숫자 인식의 실험 결과 및 분석

이 장에서는 6개의 개별 숫자 인식기 *E1, E2, E3, E4, E5, E6* 등이 포함된 인식기 풀로부터 다양한 개수의 숫자 인식기로 구성된 다수 인식기 시스템들을 구축하고, 다수 인식기 시스템의 성능 평가를 위하여 필기 숫자의 인식을 실험하고자 한다. 이들 인식기는 [13,14]에 있는 다양한 특징과 널리 알려진 특징을 사용하거나, 필기 숫자의 구조적 지식을 사용하여 KAIST와 전북대학교에서 개발되었다. 이들 인식기 중에서 일부는 신경망을 기반으로 단일 인식기 또는 클래스별 특화된 다수 인식기 결합의 형태로 구현되었고(*E1, E2, E3, E6*), 나머지는 규칙을 기반으로 클래스별로 특화된 다수 인식기 결합의 형태로 구현되었다(*E4, E5*). 이들 인식기는 단일 수준 또는 순위 수준에서 인식 결과를 결정한다.

표 1 훈련 데이타에 대한 개별 인식기의 인식 성능

인식기	훈련 방법	사용된 주요 특징	Concordia 대학 DB			UCI DB		
			정인식율	기각율	신뢰율	정인식율	기각율	신뢰율
E1	BP 신경망	픽셀 거리 함수	98.00	0.00	98.00	94.59	0.00	94.59
E2	BP 신경망	클래스별 인식기 결합, 방향 거리 분포	92.85	0.00	92.85	99.24	0.00	99.24
E3	BP 신경망	클래스별 인식기 결합, 메쉬(mesh) 특징	89.75	8.45	98.03	97.23	2.17	99.39
E4	규칙 기반	클래스별 인식기 결합, 경계, 수평 획의 굵기와 중심 등	94.08	4.65	98.67	69.26	25.97	93.56
E5	규칙 기반	클래스별 인식기 결합, 경계, 수직 및 수평 방향의 런의 수, 수평 획의 굵기 등	89.63	8.98	98.47	70.15	25.37	94.00
E6	BP 신경망	컨투어(contour) 방향 특징	96.48	0.00	96.48	99.01	0.00	99.01

표 2 테스트 데이타에 대한 개별 인식기의 인식 성능

인식기	Concordia 대학 DB			UCI DB		
	정인식율	기각율	신뢰율	정인식율	기각율	신뢰율
E1	96.00	0.00	96.00	93.77	0.00	93.77
E2	95.95	0.00	95.95	97.11	0.00	97.11
E3	84.45	12.25	96.24	91.82	5.29	96.95
E4	90.95	8.15	99.02	67.67	27.32	93.11
E5	88.15	10.40	98.38	70.01	26.15	94.80
E6	94.15	0.00	94.15	96.66	0.00	96.66

또한, 인식기의 기각 결과는 MC 선택 기준에서 사용되었으나, CE 선택 기준에서는 사용되지 않았다. Concordia 대학 CENPARMI 연구소의 필기 숫자 데이타베이스[10]에는 4000자의 훈련 데이타와 2000자의 테스트 데이타가 있으며, UCI에서 얻은 필기 숫자 데이타베이스[11]에는 3823자의 훈련 데이타와 1797자의 테스트 데이타가 있다. 개별 인식기에 대한 간략한 소개와 인식 성능은 각각 아래 표 1과 2에 나타나 있다. 인식기 *E4*와 *E5*는 Concordia 대학의 숫자 데이타를 기반으로 추출된 구조적 지식으로 구현되어 있어서, UCI의 숫자 데이타에 대한 인식 성능이 좋지 않은 편이다.

제안된 인식기의 선택 기준은 제한된 인식기의 개수에 따라서 가능한 개별 인식기 집합에 적용되어 훈련 데이타로부터 각 집합에 있는 인식기의 결과를 고려함으로써 최고의 기준치를 지닌 인식기 집합을 다수 인식

표 3 3개의 인식기로 구성된 다수 인식기 시스템

사용 DB	선택 기준	인식기 집합
UCI	FRR,RR,MCbFO,MCbCFO,MCbSO	*E2,E3,E6*
	CEbFO,CEbSO	*E2,E4,E6*
Concordia	FRR	*E1,E2,E6*
	RR	*E3,E4,E6*
	MCbFO,MCbCFO,MCbSO	*E1,E4,E6*
	CEbFO,CEbSO	*E2,E4,E6*

사용 DB	결합 방법	최고 인식기 집합
UCI	CIAB,FODB,CFODB,SODB	*E2,E3,E5*
	ΔFODB,ΔSODB	*E2,E5,E6*
Concordia	CIAB	*E1,E4,E5*
	FODB	{*E1,E4,E5*},{*E2,E3,E4*}
	CFODB,SODB	{*E1,E4,E5*},{*E1,E4,E6*}
	ΔFODB	*E2,E3,E5*
	ΔSODB	{*E2,E4,E5*},{*E2,E4,E6*}

사용 DB	결합 방법 \ 선택 기준	FRR,RR,MCbFO,MCbCFO,MCbSO	CEbFO,CEbSO	최고	평균	최저
UCI	Voting 방법	96.77	96.99	97.38	95.43	91.99
	Borda 카운트	96.44	94.82	98.33	95.51	90.98
	CIAB 방법	96.88	96.83	97.77	96.78	94.82
	FODB 방법	96.88	96.83	97.83	96.73	94.49
	CFODB 방법	96.99	97.50	97.66	96.88	94.88
	SODB 방법	97.05	97.05	97.66	96.70	94.88
	ΔFODB 방법	97.16	97.38	97.94	97.00	94.88
	ΔSODB 방법	97.33	97.77	97.89	96.99	95.21

사용 DB	결합 방법 \ 선택 기준	FRR	RR	MCbFO,MCbCFO,MCbSO	CEbFO,CEbSO	최고	평균	최저
Concordia	Voting 방법	97.30	95.75	96.50	96.90	97.30	96.07	94.80
	Borda 카운트	96.25	96.00	97.50	97.65	97.75	96.97	95.30
	CIAB 방법	96.20	96.25	96.55	96.85	97.50	96.53	95.60
	FODB 방법	96.20	96.00	96.55	96.85	97.10	96.37	95.60
	CFODB 방법	97.10	96.65	97.50	97.40	97.50	96.85	95.95
	SODB 방법	95.80	96.65	97.50	97.25	97.50	96.82	95.80
	ΔFODB 방법	97.10	96.60	97.55	97.45	97.65	97.07	96.35
	ΔSODB 방법	97.30	96.45	97.50	97.65	97.65	97.02	96.00

표 4 4개의 인식기로 구성된 다수 인식기 시스템

사용 DB	선택 기준	인식기 집합
UCI	FRR,RR,MCbFO,MCbCFO,MCbSO	*E1,E2,E3,E6*
	CEbFO	*E2,E3,E4,E6*
	CEbSO	*E3,E4,E5,E6*
Concordia	RR	*E1,E3,E4,E5*
	MCbFO,MCbCFO,MCbSO	*E1,E4,E5,E6*
	FRR,CEbFO,CEbSO	*E1,E2,E4,E6*

사용 DB	결합 방법	최고 인식기 집합
UCI	CIAB,FODB,SODB	*E2,E3,E4,E5*
	CFODB	*E2,E4,E5,E6*
	⊿FODB	{*E2,E3,E4,E5*},{*E2,E4,E5,E6*}
	⊿SODB	{*E1,E2,E3,E4*},{*E1,E2,E4,E6*}
Concordia	CIAB	*E1,E3,E4,E5*
	FODB	*E2,E3,E4,E6*
	CFODB,SODB	{*E1,E3,E4,E5*},{*E1,E3,E4,E6*}
	⊿FODB	{*E1,E3,E4,E6*},{*E1,E2,E4,E6*}
	⊿SODB	*E2,E4,E5,E6*

사용 DB	선택 기준 / 결합 방법	FRR,RR,MCbFO, MCbCFO,MCbSO	CEbFO	CEbSO	최고	평균	최저
UCI	Voting 방법	97.27	97.66	95.83	97.77	96.43	94.10
	Borda 카운트	96.66	97.44	96.83	98.05	96.58	94.38
	CIAB 방법	96.83	97.05	96.94	97.66	97.03	96.66
	FODB 방법	96.83	97.05	97.38	97.77	97.04	96.22
	CFODB 방법	97.22	97.33	97.50	98.27	97.34	96.61
	SODB 방법	97.38	97.66	97.50	97.72	97.20	95.94
	⊿FODB 방법	97.44	97.83	96.88	98.05	97.58	96.88
	⊿SODB 방법	97.38	97.50	97.44	98.05	97.60	97.16
Concordia	선택 기준 / 결합 방법	MCbFO,MCbCFO, MCbSO	RR	FRR,CEbFO, CEbSO	최고	평균	최저
	Voting 방법	97.15	97.05	97.80	97.80	97.05	96.65
	Borda 카운트	97.15	97.75	97.80	98.00	97.61	96.90
	CIAB 방법	97.00	97.45	97.05	97.45	96.83	95.70
	FODB 방법	96.55	97.00	95.85	97.15	96.53	95.70
	CFODB 방법	97.70	97.80	96.20	97.80	97.23	96.10
	SODB 방법	97.50	97.80	97.05	97.80	97.26	96.20
	⊿FODB 방법	97.90	97.65	98.00	98.00	97.72	97.15
	⊿SODB 방법	97.60	97.65	97.85	97.95	97.62	97.05

기 시스템 후보로서 선택한다. 의존관계의 차수에 따라 달라지는 정보 이론적 선택 기준을 간단히 표현하기 위하여 다음과 같은 약어를 사용한다: 즉, MCbFO는 1차 의존관계에 의한 MC 기준을 의미하고, MCbCFO는 조건부 1차 의존관계에 의한 MC 기준을, MCbSO는 2차 의존관계에 의한 MC 기준을, CEbFO는 1차 의존관계에 의한 CE 기준을, CEbSO는 2차 의존관계에 의한 CE 기준을 각각 의미한다. 그러므로, 2차 의존관계까지 고려한 정보 이론적 선택 기준의 경우에, MC 기준은 의존관계의 차수에 따라서 MCbFO, MCbCFO, MCbSO 등으로 세분화되고, CE 기준은 CEbFO와 CEbSO로 세분화된다.

3개에서 5개까지의 인식기로 구성된 다수 인식기 시스템은 테스트 데이타에 대해서 다음과 같은 결합 방법들에 의해서 평가된다: 즉, Voting 방법과 Borda 카운트 방법 외에, CIAB 방법은 조건부 독립 가정에 기반한 베이지안 방법을 나타내고, FODB 방법은 1차 의존관계에 기반한 베이지안 방법을, CFODB 방법은 조건부 1차 의존관계에 기반한 베이지안 방법, SODB 방법은 2차 의존관계에 기반한 베이지안 방법을, ⊿FODB 방법은 ⊿상호정보에 의한 1차 의존관계에 기반한 베이지안 방법을, ⊿SODB 방법은 ⊿상호정보에 의한 2차 의존관계에 기반한 베이지안 방법을 나타낸다. 이들 결합 방법은 [1,4,12]에 자세히 설명되어 있다.

표 5 5개의 인식기로 구성된 다수 인식기 시스템

사용 DB	선택 기준	인식기 집합
UCI	FRR,RR,MCbFO,MCbCFO,MCbSO	*E1,E2,E3,E5,E6*
	CEbFO	*E1,E2,E3,E4,E6*
	CEbSO	*E2,E3,E4,E5,E6*
Concordia	RR	*E1,E2,E3,E4,E5*
	FRR,CEbFO,MCbFO,MCbCFO,MCbSO	*E1,E2,E4,E5,E6*
	CEbSO	*E2,E3,E4,E5,E6*

사용 DB	결합 방법	최고 인식기 집합
UCI	CIAB,SODB	*E2,E3,E4,E5,E6*
	FODB	{*E1,E2,E3,E4,E5*},{*E2,E3,E4,E5,E6*}
	CFODB, ⊿FODB	*E1,E2,E3,E5,E6*
	⊿SODB	*E1,E2,E4,E5,E6*
Concordia	CIAB	*E2,E3,E4,E5,E6*
	FODB,SODB	*E1,E2,E3,E4,E6*
	CFODB, ⊿FODB	*E1,E3,E4,E5,E6*
	⊿SODB	{*E1,E2,E4,E5,E6*},{*E2,E3,E4,E5,E6*}

사용 DB	선택 기준 / 결합 방법	FRR,RR,MCbFO,MCbCFO,MCbSO	CEbFO	CEbSO	최고	평균	최저
UCI	Voting 방법	97.94	97.72	98.16	98.16	97.77	97.16
	Borda 카운트	97.55	97.66	97.22	97.83	97.36	96.16
	CIAB 방법	97.22	97.05	97.33	97.33	97.12	96.99
	FODB 방법	97.22	97.05	97.44	97.44	97.17	96.83
	CFODB 방법	97.89	97.61	97.61	97.89	97.65	97.50
	SODB 방법	97.77	97.44	98.27	98.27	97.61	96.88
	⊿FODB 방법	98.44	97.89	98.27	98.44	98.08	97.55
	⊿SODB 방법	98.11	98.05	98.22	98.33	98.12	97.72
Concordia	선택 기준 / 결합 방법	FRR,CEbFO,MCbFO,MCbCFO,MCbSO	RR	CEbSO	최고	평균	최저
	Voting 방법	97.90	97.40	97.45	97.90	97.55	97.40
	Borda 카운트	97.90	98.15	97.85	98.15	97.96	97.80
	CIAB 방법	97.45	97.40	97.60	97.60	97.30	96.80
	FODB 방법	97.10	96.95	96.85	97.20	96.98	96.80
	CFODB 방법	97.60	97.45	97.20	98.00	97.63	97.20
	SODB 방법	97.35	97.30	97.25	97.85	97.48	97.25
	⊿FODB 방법	98.15	97.90	98.10	98.25	98.00	97.65
	⊿SODB 방법	97.95	97.80	97.95	97.95	97.88	97.80

첫째, 3개의 인식기로 구성된 20개의 가능한 인식기 집합으로부터 제안된 선택 기준에 의하여 각기 다수 인식기 시스템 후보로 선정된 인식기 집합이 표 3의 첫 번째 테이블에 나타나 있다. 두 번째 테이블에는 실험에 사용된 다수 인식기의 결합 방법을 기준으로 실제로 최고의 인식율을 보이는 최고 인식기 집합을 나타내고 있다. 각 선택 기준에 의해서 선정된 다수 인식기 시스템 후보들의 성능은 세 번째 테이블에서 가능한 인식기 집합들에 대해서 구한 최고 인식율(최고 열), 평균 인식율(평균 열), 최저 인식율(최저 열) 등과 함께 보여지고 있다. 이 표들로부터, 두 CE 기준에 의한 다수 인식기 시스템이 다수의 경우에 있어서 다른 선택 기준에 의한 다수 인식기 시스템 보다 더 우수함을 알 수 있었다.

둘째, 4개의 인식기로 다수 인식기 시스템을 구성하는 경우에 있어서는, 15개의 가능한 인식기 집합으로부터 제안된 선택 기준에 의해 각기 다수 인식기 시스템 후보들이 선정되었으며, 이들 다수 인식기 시스템 후보들에 대한 인식 성능의 비교는 표 4에 보여지고 있다. 이 표 4에서는, CEbFO 기준에 의한 다수 인식기 시스템 후보들이 가장 우수한 다수 인식기 시스템과 일치하지 않는다 하더라도 Concordia 대학의 숫자 데이타베이스를 사용하여 FODB, CFODB, SODB 등의 결합 방법을 적용한 경우를 제외하고, 다른 선택 기준에 의한 다수 인식기 시스템 후보들 보다 더 우수한 결과를 보여 주었다.

마지막으로, 5개의 인식기로 다수 인식기 시스템을 구성하는 6개의 가능한 인식기 집합으로부터 제안된 선택

기준에 의해 각기 다수 인식기 시스템 후보들이 선정되었으며, 이들 다수 인식기 시스템 후보들에 대한 인식 성능의 비교는 표 5에 보여지고 있다. 이 표 5에서는, CEbSO 기준에 의한 다수 인식기 시스템 후보들은 UCI의 숫자 데이타베이스에 대한 실험 결과에서 우수한 결과를 보여 주었으며, 반면에 CEbFO 기준에 의한 다수 인식기 시스템 후보들은 Concordia 대학의 숫자 데이타베이스에 대해 FODB, CFODB, SODB 등의 결합 방법을 적용한 경우를 제외하고는 우수한 결과를 보여 주었다. CEbSO 기준의 경우, 흥미롭게도 두 종류의 숫자 데이타베이스에 대해 각각 여덟 가지의 결합 방법을 적용한 결과, 그 중에서 여섯 가지 경우에 가장 우수한 다수 인식기 시스템과 일치함을 보여 주었다.

이러한 실험 결과로부터, 대체로 CE 기준에 기반한 정보 이론적 선택 기준이 MC 기준에 기반한 정보 이론적 선택 기준이나 단순한 선택 기준 보다 인식기 풀로부터 성공적인 다수 인식기의 집합을 선정하는데 유용함을 알 수 있었다. 즉, 제안된 정보 이론적 선택 기준에 의해서 선정된 다수 인식기 시스템이 반드시 가장 우수한 다수 인식기 시스템과 일치하지 않는다 하더라도 우수한 인식 성능을 보여줌으로써 인식기 선택 문제에 있어서 유망한 접근 방법 중의 하나 임을 보여 주었다고 볼 수 있다. 결합 방법에 따른 인식기 선택 기준에 대해서 살펴보면, Voting 방법의 경우, CE 기준에 기반한 정보 이론적 선택 기준이 우수한 몇 가지 경우는 제외하고는 대체로 FRR 기준이 유망한 다수 인식기 시스템을 구성하였다. Borda 카운트 방법의 경우엔 특별히 강한 영향을 미치는 인식기 선택 기준이 없었다. 끝으로, 의존관계에 기반한 베이지안 결합 방법을 사용하는 경우엔 CE 기준에 기반한 정보 이론적 선택 기준이 다른 선택 기준 보다 유망한 다수 인식기 시스템을 구성했다고 볼 수 있다. 이상과 같은 실험 결과로부터, 제안된 정보 이론적 선택 기준이 인식기 풀로부터의 인식기 선택 문제에 대한 유망한 접근 방법중의 하나라고 여길 수 있다.

5. 결 론

이상과 같이 인식기 풀로부터 다수 인식기를 선택하는데 있어서 다수 인식기를 결합하는 방법에 대한 고려가 필요하다고 볼 수 있다. 단순한 선택 기준과 두 종류의 정보 이론적 선택 기준을 두 종류의 필기 숫자 데이타베이스에 적용하여 다수 인식기 시스템을 구성하였고, 각 선택 기준에 따른 성능 평가도 수행하였다. 이러한 필기 숫자 인식 실험을 통하여 제안된 정보 이론적 선택 기준이 다수 인식기 시스템의 구성을 위한 긍정적인 결과를 보여 주었다 하더라도 더 깊이 있는 연구가 필요하다. 왜냐하면, 선정된 다수 인식기 시스템이 종종 가장 우수한 인식 결과를 보장하지 않을 뿐만 아니라, 다수 인식기 시스템에 포함될 인식기 개수가 미리 제한되어 있다는 가정이 본 연구의 제한점이기 때문이다. 게다가 [3]에서 언급하고 4장의 실험 결과에서 보듯이, 3개 인식기의 집합이 결합 방법에 따라서 4개나 5개 인식기의 집합 보다 가끔은 더 우수한 결과를 보여주는 경우가 있기 때문이다. 따라서, 불필요하거나 중복된 인식기를 선별하여 제거하는 체계적인 방법과 가장 우수한 다수 인식기 시스템을 구축하기 위한 인식기의 최소한의 집합을 알아내기 위한 향후 연구에도 많은 관심과 주의가 필요하다고 본다.

참 고 문 헌

[1] Xu, L., Krzyzak, A., and Suen, C. Y., "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," *IEEE Trans. on Systems, Man, and Cybernetics*, 22(3):418-435, 1992.

[2] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J., "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226-239, 1998.

[3] Woods, K., Kegelmeyer Jr., W. P., and Bowyer, K., "Combinition of Multiple Classifiers Using Local Accuracy Estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):405-410, 1997.

[4] 강희중, 이성환, "무제약 필기 숫자를 인식하기 위한 다수 인식기를 결합하는 의존관계 기반의 프레임워크", 정보과학회논문지 : 소프트웨어 및 응용, 제27권, 제8호, pp. 855-863, 2000.

[5] Kim, J., Seo, K., and Chung, K., "A Systematic Approach to Classifier Selection on Combining Multiple Classifiers for Handwritten Digit Recognition," In *Proceedings of the 4th International Conference on Document Analysis and Recognition*, vol. 2, pp. 459-462, 1997.

[6] Impedovo, S. and Salzo, A., "Evaluation of Combination Methods," In *Proceedings of the 5th International Conference on Document Analysis and Recognition*, pp. 394-397, 1999.

[7] Lewis, P. M., "Approximating Probability Distributions to Reduce Storage Requirement," *Information and Control*, 2:214-225, Sep. 1959.

[8] Gallager, R. G., *Information Theory and Reliable Communication,* John Wiley and Sons, Inc., 1968.
[9] Wang, D. C. C. and Wong, A. K. C., "Classification of Discrete Data with Feature Space Transform," *IEEE Transactions on Automatic Control*, AC-24(3):434-437, 1979.
[10] Suen, C. Y., Nadal, C., Legault, R., Mai, T. A., and Lam, L., "Computer Recognition of Unconstrained Handwritten Numerals," In *Proceedings of IEEE*, pp. 1162-1180, 1992.
[11] Blake, C. and Merz, C., UCI repository of machine learning databases, 1998.
[12] 강회중, "베이스 에러율의 상위 경계 최소화에 기반한 고차 곱 근사 방법과 숫자 인식기 결합에의 적용", 정보과학회논문지 : 소프트웨어 및 응용, 제28권, 제9호, pp. 681-687, 2001.
[13] Matsui, T., Tsutsumida, T., and Srihari, S. N., "Combination of Stroke/Background Structure and Contour-direction Features in Handprinted Alphanumeric Recognition," In *Proceedings of the 4th Int. Workshop on Frontiers in Handwriting Recognition*, pp. 87-96, 1994.
[14] Oh, I.-S. and Suen, C. Y., "Distance features for neural network-based recognition of handwritten characters," *International Journal on Document Analysis and Recognition*, 1(2):73-88, 1998.

강 희 중

1986년 서울대학교 전자계산기공학과 졸업(학사). 1988년 한국과학기술원 전산학과 졸업(석사). 1997년 한국과학기술원 전산학과 졸업(박사). 1986년 ~ 1998년 삼성전자주식회사 기업통신개발그룹 선임연구원. 1998년 ~ 2000년 고려대학교 인공시각연구센터 연구조교수. 2000년 ~ 현재 한성대학교 정보전산학부 교수. 관심분야는 인공지능, 패턴인식, 그룹웨어