韓國 컴퓨터情報學會 論文誌
第12卷 第6號, 2007. 12.

2007-12-6-1-20

GPCR 분류에서 ART1 군집화를 위한 퍼지기반 임계값 제어 기법

조 규 철*, 마 용 범*, 이 종 식*

Fuzzy-based Threshold Controlling Method for ART1 Clustering in GPCR Classification

Cho Kyu Cheol*, Ma Yong Beom*, Lee Jong Sik*

요 약

퍼지이론은 생명정보공학에서 지식을 표현하는데 활용되고 제어시스템 모델을 이해하는데 활용되어 왔다. 본 논문에서는 생명정보학의 응용 프로그램에서 중요한 데이터 분류에 초점을 맞추었다. 최적의 임계값 유도를 위한 GPCR 분류에서 기존의 순차기반 임계값 제어기법은 임계값 결정범위와 최적의 임계값 유도 시간의 문제점을 보였고, 이진기반 임계값 제어기법은 임계값 결정 초기에 시스템의 안정성에 대한 단점이 있었다. 이를 보완하기 위해 우리는 ART1 군집화를 위한 퍼지기반 임계값 제어기법을 제안한다. 제안된 방법의 성능을 평가하기 위해 ART1 군집화를 위한 퍼지기반 임계값 제어기법을 구현하여 기존의 순차기반 임계값 제어기법과 이진기반 임계값 제어기법과의 인식률에 대한 구동시간의 변화, 임계값의 변화에 따른 시스템의 구동시간을 측정하였다. 퍼지기반 임계값 제어 기법은 GPCR 데이터 분류에서 인식률과 구동시간에 대한 정보를 통해 분류 임계값을 조정하여 높은 인식률과 낮은 구동시간을 지속적으로 유도하여 안정적이고 효과적인 분류 시스템을 만들 수 있었다.

Abstract

Fuzzy logic is used to represent qualitative knowledge and provides interpretability to a controlling system model in bioinformatics. This paper focuses on a bioinformatics data classification which is an important bioinformatics application. This paper reviews the two traditional controlling system models. The sequence-based threshold controller have problems of optimal range decision for threshold readjustment and long processing time for optimal threshold induction. And the binary-based threshold controller does not guarantee for early system stability in the GPCR data classification for optimal threshold induction. To solve these problems, we proposes a fuzzy-based threshold controller for ART1 clustering in GPCR classification. We implement the proposed method and measure processing time by changing an induction recognition

• 제1저자 : 조규철
• 접수일 : 2007.12. 6, 심사일 : 2007.12.15, 심사완료일 : 2007.12.17.
* 인하대학교 컴퓨터 정보공학과

success rate and a classification threshold value. And, we compares the proposed method with the sequence-based threshold controller and the binary-based threshold controller. The fuzzy-based threshold controller continuously readjusts threshold values with membership function of the previous recognition success rate. The fuzzy-based threshold controller keeps system stability and improves classification system efficiency in GPCR classification.

▸ Keyword : 퍼지(Fuzzy), GPCR, 데이터 군집화 분류, 뎁스 시뮬레이션(DEVS Simulation)

Ⅰ. 서 론

패턴인식 기술[1]은 데이터에 대한 패턴 정보를 추출하여 응용하는 기술로써 오랜 기간의 연구가 진행되고 그 결과로 실생활에서 이미 널리 사용되고 있다. 패턴인식은 인간의 감각기관을 통하여 인지된 정보를 분류, 특징 추출, 여과, 강조 등의 과정을 통하여 관찰하며 이해된 정보를 전달하는 형태로 컴퓨터를 통해 인공지능 기술로 활용한다.

이러한 패턴인식 기술 중의 퍼지이론[2]은 확률 이론과 다르게 불분명한 상태나 모호한 상태를 이진 논리에서 벗어나 정할 수 없는 수치를 표현하는 개념으로 다소 비수학적인 이론으로 뉴럴 네트워크에 활용되며 생물정보과학에서 연구자들에게 유용한 정보를 제공해준다. 특히 퍼지이론을 이용하여 데이터 분류나 인식에서 중요하게 작용하는 요소들을 제어하여 시스템의 성능과 효율을 향상시키는데 적용되고 있다[3].

모든 생명체의 생리작용에서 세포 외에서 세포 내로 정보를 전달하는 신호 전달은 가장 기본적인 생물학적 현상이다. 신호 전달은 세포나 조직의 성장, 분화, 사멸에 관련하여 긴밀한 작용을 하고, 유전자의 발현, 신경전달 물질이나 호르몬의 분비, 세포막을 통한 이온의 수송 등의 미세현상의 조절을 관장하고 있다. 생물정보과학에서 이러한 신호 전달에 관한 연구에서 크게 비중을 두고 연구하고 있는 분야가 GPCR(G-protein coupled receptor)[4]에 관한 연구이다.

GPCR 데이터는 수많은 단백질 정보를 이용하여 인간 게놈 프로젝트에 활용되고 있고, 정확하고 시퀀스 데이터에 기초한 GPCR 데이터를 분류하고 예측하는 방법은 생물학적인 연구를 통해 실용적인 가치고 사용된다. 이러한 GPCR 데이터에 대해 다양한 자동 인식 기법들의 장점이 토론되면서 넓은 범위의 유기체 안에서 발견되고 생리학상의 진행을 조절하는 세포질의 신호 네트워크의 중심에서 주요한 기능을 하기 때문에 현대 약물 조제 연구[5]가 관심 받고 있고, 꾸준히 늘어나고 있는 단백질 데이터에 대한 인식이 필요하기 때문에 정확하고 자동적으로 구분할 수 있는 인식기에 대한 요구는 꾸준히 증가하고 있다.

군집화 분류에서 임계값에 대한 범위를 결정하는 방법 중 순차 기반 임계값(Sequence-based Threshold) 제어기법[6]은 결정 범위의 간격 결정이 어렵고 최적값 유도 검색도 비효율적인 단점이 있고, 이진 기반 임계값(Binary-based Threshold) 제어기법[7]은 임계값 범위 조절범위가 크기 때문에 인식률 및 빠른 구동시간이 보장되지 않는 단점이 있다.

이러한 단점을 보완하기 위해 본 논문에서는 ART1을 이용한 GPCR 분류에서 군집화의 기준이 되는 분류 임계값을 퍼지이론에 최적의 인식률과 구동시간을 제공하는 임계값을 검색하는데 이전의 인식률과 구동시간을 멤버로 활용하여 성능을 향상시키는 제어 기법을 제안하여 순차 기반 임계값 제어 기법과 이진 기반 임계값 제어 기법을 인식률에 따른 구동시간 변화와 임계값 변화 횟수에 따른 구동 시간 변화를 비교하였다.

제안하는 분류 기법에 데이터를 사용하기 위해 문자열로 구성되어 계산이 불가능한 GPCR 데이터를 조작 가능한 데이터로 전처리하였고, ART1 네트워크를 이용한 무감독 방법을 통하여 분류를 진행하였고, 이때 ART1 분류기의 구동시간과 인식률을 고려한 분류 임계값을 정하기 위하여 퍼지이론을 이용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구의 배경이 되는 생명정보학에서의 퍼지 이론과 기존의 임계값 제어기법을 기술하고 3장에서는 ART1 군집화를 위한 퍼지기반 임계값 제어기법의 방법에 대해서 기술한다. 4장에서는 GPCR 데이터를 기술하고 ART1 데이터 분류에서 퍼지기반 분류 임계값 제어기법의 구현하며 실험을 통한 제안 기법의 안정성과 유용함을 입증한다. 그리고 5장에서는 결론을 맺는다.

Ⅱ. 관련연구

2.1 생명정보학에서 퍼지이론

생명정보학에서 퍼지이론을 적용한 인공지능 기반의 분류는 여러 문제에 적용하여 사용되고 있다. 계층적 퍼지 분류기는 퍼지 규칙 기반의 분류기에서 규칙을 확장하여 문제를 잠

재적으로 완화시켜 이를 활용하여 단백질의 하부 구조 예측에서 과거의 분류된 기록들보다 더 뛰어나게 두 번째 레벨의 하부 계층을 예측하였다[8].

퍼지 HMM(Hidden Markov Model)[9]은 음성인식에 적용되어 음성의 외형 HMM 표현 방법은 퍼지 적분과 퍼지 오퍼레이터들로 정의되어 퍼지 가능성에 대체하여 HMM 확률로 정의하여 음성을 판단하였다. 그리고 FGMM(Fuzzy Gaaussian Mixure Model)은 효과적인 군집 알고리즘으로 음성인식과 이미지 인식에 적용되어 왔다[10].

뉴럴 퍼지는 다중 계층 네트워크에서 퍼지 언어학 알고리즘으로 설계되고 뉴런의 결합 구조를 통해 퍼지 추론과 적응의 진행을 정확하게 반복하기 위해 구현한다[11].

최근에는 SVM(Support Vector Machines)을 진보된 퍼지이론과 결합하여 데이터 분류에 더 진보된 분류를 가능하게 하였다[12].

2.2 기존의 임계값 제어기법

2.2.1 순차 기반 임계값 제어기법

순차 기반 임계값 제어기법은 수식(1)과 같이 순차 기반의 결정범위를 제어[6]하여 일정한 간격으로 임계값을 증가시키며 제어하는 기법으로, 임계값 조정 간격을 적게하는 경우 인식률을 세밀하게 조절하며 원하는 인식률을 유도가 가능하지만 그만큼 시도횟수가 많아지기 때문에 구동시간에 비효율적이다. 반면에 임계값 조정 간격을 크게하는 경우 시도횟수가 적어 구동시간은 적게 걸리지만, 인식률을 세밀하고 조절할 수 없을 뿐 아니라 원하는 인식률을 검색하지 못하는 경우가 발생하여 비효율적인 문제가 발생할 수 있다.

$$T_{n+1} = T_n + \alpha \text{(단, } T_0 = Min(T)) \quad \cdots\cdots (1)$$

수식(1)에서 T는 경계값을 의미하며 α는 임계값 조정 간격을 의미한다. 그리고 T0는 임계값의 초기값으로 ART1 군집화를 위한 초기값은 1이다.

순차 기반 임계값 제어기법은 유도하는 최적 임계값의 검색이 늦어지면 구동시간의 효율을 기대할 수 없고 임계값 결정 범위가 크면 원하는 임계값 유도 여부도 보장할 수 없는 단점이 있다.

2.2.2 이진 기반 임계값 제어기법

이진 기반 임계값 제어 기법은 수식(2)와 같이 이진 기반의 결정범위를 제어[7]하여 임계값의 최소값과 최대값을 기준으로 임계값 조정 결정 범위를 정한 후, 이전의 임계값에서 인식률이 낮출 경우 임계값을 임계값 조정 결정 범위를 빼주어, 인식률을 높일 경우 임계값을 임계값 조정 결정 범위를 더해주어 인식률을 높이는 기법이다. 이때 임계값 조정 결정 범위는 이전의 임계값 조정 결정 범위를 2로 나누어 조정하게 된다.

$$\beta_{n+1} = \frac{\beta_n}{2} \text{ (단, } \beta_0 = Max(T) - Min(T)) \quad \cdots\cdots (2)$$
$$T_{n+1} = T_n \pm \beta_n \text{ (단, } T_0 = Min(T))$$

수식(2)에서 β는 임계값 조정 간격을 의미한다. β0는 임계값의 최대값을 의미하는 Max(T)와 최소값을 의미하는 Mix(T)의 간격을 의미하고, ART1 군집화를 위한 초기값은 1이다.

이 기법은 초기에 임계값 조정 결정 범위가 크기 때문에 인식률이 낮거나 빠른 구동시간을 보장할 수 없는 단점이 있다. 특히 임계값을 올릴 것인지 내릴 것인지 결정을 한번 잘못하면 계속 문제가 발생할 수 있다.

III. ART1 군집화를 위한 퍼지기반 임계값 제어 기법

본 논문에서는 GPCR 데이터 분류에서 ART1 군집화를 위한 퍼지기반 임계값 제어 기법을 제안한다. 여기서는 ART1 분류 네트워크에 대한 소개와 ART1 군집화에서 사용되는 임계값에 대한 설명과 임계값을 조정하기 위해 적용된 퍼지기반 임계값 제어 기법에 대해서 기술한다.

3.1 ART1 분류 네트워크

ART1(Adaptive Resonance Theory 1)[13] 분류 네트워크는 데이터 경쟁 학습의 약점이었던 안정성을 보강하여 제안한 신경회로망 모델로써 뉴럴 네트워크의 안정성과 적응성 문제를 해결하였고, 특히 무감독 분류 기법을 특징으로 기존에 학습되던 것에 대하여 새로운 학습에 의해 지워지지 않고 자동적으로 기존 전체 지식 베이스에 일관성 있게 통합하여 훈련된다.

ART1은 반복적인 학습을 통하여 적절하게 매치되는 새로운 정보는 기존에 배운 내용들을 정제하여 갱신하며, 인식 카테고리의 학습을 위해 새로운 유닛을 선택하여 새로운 대표 군집화를 생성하고, 기억용량의 한계를 넘어서는 새로운 입력에 의해 기존에 취득한 내용이 지워지지 않는다. 그렇기 때문에 지속적인 데이터 입력이나 제한 없는 입력에 대해서 자신의 메모리 용량을 전부 소모할 때까지는 동적으로 실시간으로

빠르고 안정되게 학습할 수 있는 네트워크 구조이다. 이를 통해 구성된 결과는 생명정보학에서 단백질의 구조와 기능을 밝혀내고 새로운 단백질 카테고리를 구성하는 데에 있어서도 중요한 역할을 할 수 있다.

그러나 무감독 분류를 통해 임계값과 비교하며 새로운 대표 군집이 자동 생성되어 저장되기 때문에 저장방법이 매우 비효율일 수 있으며, 특히 많은 수의 대표 패턴들을 저장할 수가 없기 때문에 대표 패턴에 제한이 있는 경우가 많다.

3.2 ART1 군집화에서의 임계값

ART1을 이용한 분류는 무감독 분류학습에 속한다. 이는 무감독 학습법에 의해 기존의 군집들과 비교하여 유사한 군집이 없는 경우 새로운 군집을 생성한다. 이때 비교하는 기준이 임계값이 되는데, 임계값을 어떻게 정하느냐에 따라 군집의 수가 정해진다.

GPCR 분류를 위한 ART1 군집화에서는 임계값 기준 척도가 높아지는 경우 클러스터 유사도가 높아야 하기 때문에 군집들이 많이 생성된다. 이 경우는 비교 대상이 많아지기 때문에 많은 군집속에서 새로운 데이터와 가장 유사한 패턴을 검색할 수 있어 인식률이 좋아진다는 장점이 있으나 생성된 군집들과 유사성 측정을 모두 해주어야 하기 때문에 구동시간은 군집화의 수에 비례하여 증가한다. 그리고 군집들이 많아지면서 해당 군집들의 패턴에 대한 저장에 대한 효율성 및 관리에 대한 문제도 충분히 고려해야한다. 반면에 임계값이 낮은 경우 군집이 적게 생성되어 구동시간은 절약할 수 있으나 유사한 패턴 검색에 실패하는 경우가 빈번하게 발생하여 인식률 저하를 초래할 수 있다.

ART1 군집화를 이용한 데이터 분류는 인식률과 구동시간을 고려하여 최적의 인식률과 해당 인식률을 제공하며 시간을 절약할 수 있는 임계값을 정해야한다. 본 논문에서는 임계값을 정하기 위해 퍼지이론을 이용하여 데이터 분류를 진행하며 인식률과 구동시간을 비교 판단하여 임계값을 갱신한다.

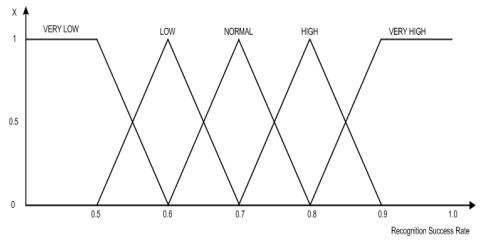
3.3 퍼지기반 임계값 제어

ART1 군집화를 위한 퍼지기반 임계값 제어는 맘다니 모델[14]을 이용하여 추론하고 역퍼지화를 위해 무게 중심법[15]을 사용하였다.

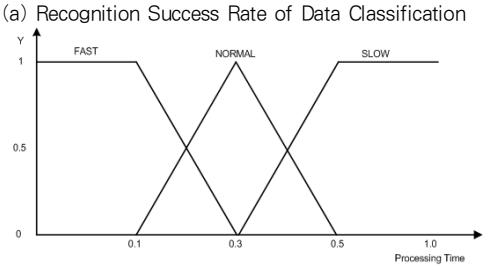
임계값 제어를 위한 퍼지 로직의 입력 파라미터는 GPCR 데이터 분류에서의 인식률(X)과 데이터 분류기의 구동시간(Y)이며, 퍼지 로직의 출력 파라미터는 분류 경계값(Z)이다.

X(인식성공률) = { VERY LOW, LOW, NORMAL, HIGH, VERY HIGH }
Y (구동시간) = { FAST, NORMAL, SLOW }
Z (분류 경계값) = { LOW, NORMAL, HIGH }

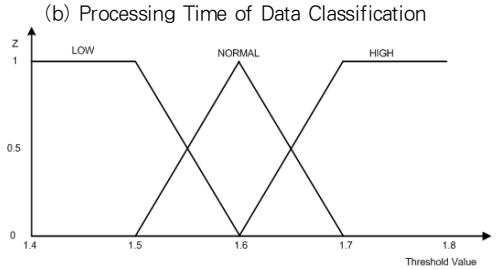
그림 1은 ART1 군집화를 위한 임계값을 갱신하기 위해 15가지의 퍼지 규칙을 정하여 분류 인식률과 데이터 분류기의 구동시간에 관련한 입력 멤버십 함수와 3개의 멤버십을 가지는 임계값을 정하기 위한 출력 멤버십 함수를 나타낸 것이다.



(a) 분류 인식률
(a) Recognition Success Rate of Data Classification



(b) 분류 구동시간
(b) Processing Time of Data Classification



(c) 경계 임계값
(c) Threshold Value
그림 1. 입/출력 파라미터 멤버십 함수
Fig. 1 Fuzzy Sets of Input and Output

ART1 분류가 진행되면서 데이터의 인식률과 구동시간이 출력되면 기록된 두 개의 값을 기반으로 출력 소속함수를 자신이 해당하는 소속값에서 잘라낼 수 있으며 그림 2의 퍼지 규칙을 통해서 찾아 볼 수 있다.

```
Rule 0: IF ( X is VERY LOW ) AND ( Y is SLOW )
        THEN ( Z is HIGH )
Rule 1: IF ( X is VERY LOW ) AND ( Y is NORMAL )
        THEN ( Z is HIGH )
Rule 2: IF ( X is VERY LOW ) AND ( Y is FAST )
        THEN ( Z is NORMAL )
Rule 3: IF ( X is LOW ) AND ( Y is SLOW )
        THEN ( Z is HIGH )
        ...................
        ...................
Rule 15: IF ( X is VERY HIGH ) AND ( Y is FAST )
        THEN ( Z is NORMAL )
```

그림 2. 임계값 제어를 위한 퍼지 규칙
Fig. 2 Fuzzy rules for threshold value control

수식(3)은 분류 인식률과 분류기의 구동시간의 값으로 퍼지 이론에 적용하여 다음 경계 임계값을 책정하기 위한 수식이다.

$$T = \frac{(RSR_x \bullet M_x) + (PT_y \bullet M_y)}{RSR_x + PT_y} \times Max(T) \quad \cdots\cdots\cdots\cdots (3)$$

수식에서 RSRx, PTy는 입력 X와 Y값에 대한 멤버십 값이고 Mx와 My는 RSRx, PTy의 출력 Z에 대칭되는 값이다. 그리고 Max(T)는 경계 임계값이 가질 수 있는 최대의 값이다.

Ⅳ. 실험 및 결과

본 논문에서는 GPCR 분류에서 ART1 군집화를 위한 퍼지 기반 임계값 제어 기법을 활용하여 퍼지 이론을 적용하여 향상된 성능을 알아보기 위해 인식률의 변화 과정과 구동시간을 통해 알아본다. GPCR 데이터에 대한 정보와 이를 퍼지이론을 장착한 ART1 군집화 분류기에 데이터를 적용하기 위해 데이터에 대한 전처리과정과 구현방법에 대해서 기술한다. 그리고 실험을 통해 제안한 기법의 유용함과 성능을 입증한다.

4.1 실험 데이터: GPCR 데이터 및 전처리 과정

GPCR 데이터는 Class A, Class B, Class C, Class D, Class E와 decoy로 크게 상위 클래스로 나눌 수 있고, GPCR 데이터 정보 시스템에 따르면 각 클래스는 692, 56, 16, 11, 3, 99의 데이터로 구성되어 있다[16]. 본 논문에서는 5개 클래스로 구성된 GPCR을 Class A와 나머지 클래스들을 The Others 클래스로 대분류하여 데이터를 구성하였다.

생명정보학에서 GPCR은 데이터들을 클래스화하여 구분하기 위하여 SVM은 보이지 않는 표본의 분류 자격에 대해서 정확한 예측을 위해 커널 함수라는 수학적 툴을 사용하여 두 분류 사이에 유사성을 측정하여 같은 배위자를 묶은 것처럼 small subfamily인식에 사용되어 낮은 에러율과 메소드를 지원하여 정확하게 구별[17]하고 있고, 자동적으로 데이터 분류를 위하여 HMM를 통하여 데이터를 정확하게 구분하고 있고[18]. 그리고 FSV(Fisher score vectors)[19]에 매핑하여 단백질 분류를 위한 SVM library를 만들고 전체 library에 관한 결과 값으로 분류하였다. 다중 클래스 분류 중에 Class A 와 Class C의 Subfamily 구별을 위한 기계학습 커뮤니티와 다양한 학습 문제들에 대해서 extended binary SVM[20]을 사용하여 두개 이상의 데이터 셋을 훈련하는 경우 하나의 클래스를 positive class로 라벨링하였고 나머지 클래스를 negative class로 라벨링하여 1:N 다중 클래스를 구별하였다.

이러한 GPCR 데이터는 각각 다른 길이의 문자열로 구성되어 있기 때문에 ART1 이상치 감시 기반 군집화 분류 기법이 고정 길이 레이어로 되어 있어 GPCR 데이터들을 바로 분류를 할 수 없기 때문에 이 문제를 해결하기 위해 데이터들 모두 같은 길이로 가공해야 하는 문제점이 있다.

이를 해결하기 위해 본 논문에서는 GPCR 데이터들을 ClustalX[21] 프로그램을 사용하여 고정 길이로 변환하여 고정 길이 입력 레이어에 맞추었다. 그림 3은 ClustalX를 통하여 GPCR 데이터를 다중 정렬과정을 표현하였다. ClustalX는 다중 서열 정렬 프로그램으로 DNA나 단백질 전역의 다중 정렬을 통해 생물학적인 의미를 찾는데 유용한 프로그램으로 대상 서열들 간의 유사성을 비교하여 연관성이 있게 정렬하여 일치, 유사, 차이점을 알아 볼 수 있어 GPCR 다중 정렬에 적용이 가능하다.

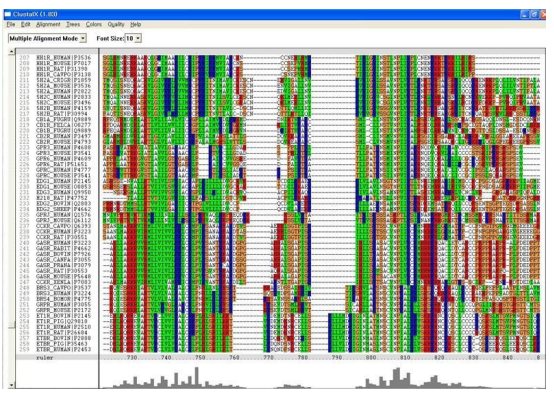


그림 3. ClustalX를 이용한 GPCR 데이터 다중 정렬
Fig. 3 GPCR Multiple Alignment Using the ClustalX

그리고 데이터를 재조정하기 위해 다중 정렬된 각 데이터에 대하여 [-]가 50%이상 포함되어 있는 구간의 데이터를 여과하여 351 길이의 데이터를 얻게 되었다. 그리고 각 문자

데이터를 0과 1사이의 실수 데이터로 맵핑하여 ART1 군집화 분류 과정에서 계산이 가능한 데이터를 얻어낼 수 있다.

4.2 실험을 위한 ART1 분류기 구현

본 논문에서는 ART1 군집화를 위한 퍼지 기반 임계값 제어 기법의 유용함과 성능을 평가하기 위해 DEVS(Discrete Event System Specification) 모델링과 시뮬레이션 환경[22][23]에서 설계하였다.

DEVS는 메시지 단위의 Event발생에 의한 방식으로 구동되며, 계층적이고 모듈화된 이산 사건의 모델들을 위해 정의된 이론으로 메시지 단위의 EVENT 발생에 의한 방식으로 구동되며 계층적이고 모듈화된 형식을 기술하여 시간의 흐름에 따라 시스템의 입력, 상태, 출력, 상태전이를 추상화한다.

그림 4는 실험을 위해 구현된 퍼지 이론을 이용하여 경계 임계값을 갱신하는 ART1 분류기의 아키텍처이다. Classifier Manager에서 ART1 군집화를 위한 경계 임계값을 정하여 Data Training Machine들에게 임계값을 전달하면 각 Data Training Machine들은 전처리된 GPCR데이터들을 이용하여 군집들을 생성하며 데이터들을 분류하고 분류 결과 및 구동시간을 Tester에게 전달하면 Tester는 테스트 데이터를 이용하여 인식률을 측정하게 되고 이에 대한 결과를 Classifier Manager에게 인식률과 구동시간 결과를 전송한다. 수신된 결과에 대하여 Classifier Manager는 인식률과 구동시간을 퍼지이론을 이용하여 다음 경계 임계값을 유도함으로써 계산시간과 인식률을 고려한 임계값을 얻어낸다.

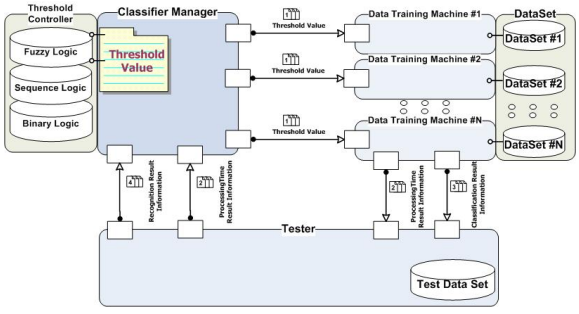


그림 4. ART1 군집화 분류기의 아키텍처
Fig. 4 Architecture of the ART1 Clustering Classifier

본 논문에서는 ART1 군집화를 위한 퍼지 기반 임계값 제어기법에서 ART1 군집화의 퍼지기반 임계값 제어기법의 유용함을 알아보기 위해서 순차 기반 임계값 제어기법과 이진 기반 임계값 제어기법을 비교하여 인식률에 따른 구동시간과 임계값 갱신에 따른 전체 구동시간의 변화를 비교하였다.

4.3 실험 결과

4.3.1 실험의 목적

ART1 군집화를 통한 분류는 군집이 많으면 유사한 패턴을 검색하기위해 비교해야 할 대상이 많아지기 때문에 유사 패턴을 찾아내는 인식률은 대부분 상승하지만 모든 군집들과 비교해야하기 때문에 구동시간은 군집의 수와 비례하여 증가하는 반면에 군집이 적으면 비교해야하는 유사 패턴이 적기 때문에 인식률은 보장받지 못한다. 그렇기 때문에 인식률에 대한 안정성을 보장하고 빠른 구동시간을 유도하여 효과적인 시스템을 만들기 위해 적은 수의 군집을 생성하는 임계값 검색을 가능하게 하는 방법을 찾아야 한다.

4.3.2 실험 결과1:인식률에 따른 구동시간 변화

그림 5는 인식률에 따른 구동시간의 변화를 측정하여 도식화한 것으로 순차 기반 임계값 제어기법은 인식률이 80, 85%를 보장하는 임계값을 4초 이내에 유도하지만 이후 90% 이상은 임계값이 계속 높아지면서 구동시간이 길어진다. 이진 기반 임계값 제어기법은 초기에 결정범위가 커서 인식률이 불안정하기 때문에 인식률을 보장하는 임계값을 유도하는 구동시간이 전반적으로 길고, 이후에는 다른 인식률 유도가 빠르고 안정한 것을 알 수 있다. 하지만 퍼지 기반 임계값 제어기법보다는 대체적으로 많은 시간을 소요했음을 알 수 있다. 그리고 퍼지 기반 임계값 제어기법은 임계값을 인식률과는 상관없이 4초에서 6초의 구동시간을 소요하며, 비슷한 시간 내에 임계값을 유도하였고 80, 85%는 순차 기반 임계값 제어기법보다 오랜 시간이 걸리지만 90, 95, 98%의 경우에는 다른 기법들보다 최적의 인식률 유도에 효과가 있음을 보여주고 있다.

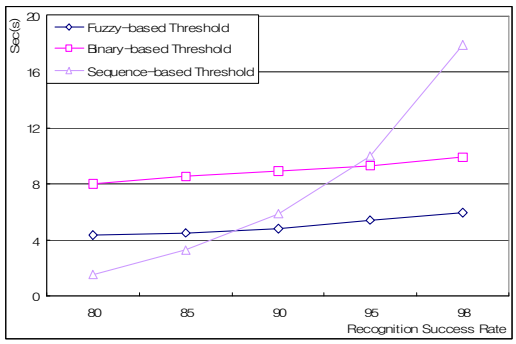


그림 5. 인식률에 따른 구동시간 변화
Fig. 5 Processing Time with Recognition Sucess Rate

4.3.3 실험 결과 2: 임계값 변화 횟수에 따른 구동 시간 변화

그림 6은 임계값 갱신에 따른 구동 시간의 변화를 측정하여 도식화한 것이다. 분류 임계값이 낮은 경우에는 생성되는 클러스터가 적기 때문에 구동시간이 적게 소요되지만 임계값이 높아지는 경우 생성되는 클러스터가 많아져서 구동시간이 오래 걸리게 된다. 순차 기반 임계값 제어기법은 초기에 분류 임계값이 낮기 때문에 생성되는 군집의 수가 적어 낮은 구동시간을 기록하지만 후기에는 분류 임계값이 높아져 생성되는 군집의 수도 많아지고 이에 따라 구동시간이 길어져서 비효율적인 성능을 보이고, 이진 기반 임계값 제어기법은 초기의 이진 임계값 갱신 구간에서 높은 구동시간을 나타내지만 4번의 임계값 갱신 이후에는 안정된 구동시간을 나타내었다. 그리고 퍼지 기반 임계값 제어기법은 낮은 증가율을 제시하며 다른 두 제어기법보다 낮은 구동시간을 기록하였다.

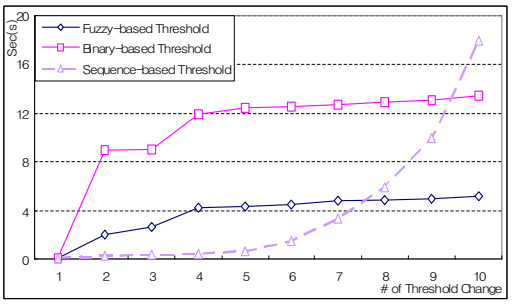


그림 6. 임계값 변화 횟수에 따른 구동 시간 변화
Fig. 6 Processing Time with Number of Threshold Value Change

4.3.4 실험의 의미

4.3.2의 그림5를 통해 퍼지 기반 임계값 제어 기법이 분류 인식률과 구동시간에 대한 멤버십 함수와 퍼지규칙을 통해 유도하는 최적의 인식률을 찾아내며 가장 빠른 시간에 최적의 시스템을 유도하는 효과가 있음을 의미하고, 4.3.3의 그림6을 통해 시스템 초기부터 구동시간을 고려하여 임계값 수정하여 낮은 구동시간을 기록하게 하고 이후에도 지속적으로 시간에 대한 효율을 높인 것을 확인할 수 있다.

두 가지 실험 결과를 통해 원하는 인식률을 유도하는 시간과 임계값 변화에 따른 시스템의 구동시간을 비교를 통해 퍼지 기반 임계값 제어기법이 순차 기반 임계값 제어기법과 이진 기반 임계값 제어기법보다 빠르게 원하는 인식률을 유도하고 구동시간도 향상된 것을 알 수 있었다. 이를 통해 ART1 군집화를 위한 퍼지기반 임계값 제어기법은 최적의 인식률과 낮은 구동시간을 빠른 시간에 제어하고 지속적으로 유지하여 안정적인 성능을 보장하는 효과가 있음을 보여준다.

V. 결론 및 향후 과제

생명정보학에서 단백질 데이터에 대한 연구의 중요성과 더불어 꾸준히 증가하는 새로운 단백질 데이터에 대한 인식 방법에 대하여 연구가 꾸준히 진행되고 있다. 본 논문에서는 단백질 데이터 중에서 안정적이고 효과적인 GPCR 데이터 분류를 위하여 ART1 군집화를 위한 퍼지기반 임계값 제어기법을 제안하였다.

우리는 GPCR의 데이터들을 전처리하여 Class A 클래스와 나머지 클래스들을 The Others 클래스로 나누었고, 퍼지 기반 임계값 제어기법을 적용한 ART1 군집화 분류기를 구현하여 기존의 순차 기반 임계값 제어기법과 이진 기반 임계값 제어기법을 임계값 갱신에 따른 임계값 변화, 인식률과 구동시간을 측정하여 성능을 비교하였다.

ART1 군집화를 통한 데이터 분류는 군집의 수에 따라 인식률과 구동시간이 비례하는데 적절한 분류 임계값을 조정하여 높은 인식률을 유지하면서 구동시간을 줄이는데 이를 위해서 퍼지기반 임계값 제어기법은 인식률과 구동시간을 비교하여 임계값을 조정함으로써 높은 인식률을 유지하며 효율적인 임계값 관리가 가능했다.

실험에서는 ART1 군집화위한 퍼지기반 임계값 제어기법은 임계값 결정 간격에 대한 안정성과 구동시간에 단점이 있었던 기존의 순차기반 임계값 제어기법과 임계값 초기 결정단계에서 불안정하였던 기존의 이진기반 임계값 제어기법과 비교하여 보다 빠른 시간에 유도하여 원하는 최적의 인식률유도에 효과가 있음을 증명하였고, 구동시간에 대한 비교에서도 다른 두 제어기법은 많은 구동시간이 요구되는데 비해 퍼지기반 제어기법은 초기부터 구동시간을 고려하여 인식률을 높여주기 때문에 유도하려는 인식률과 관계없이 구동시간이 비슷하게 소요되었다.

본 논문에서 제안한 ART1 군집화를 위한 퍼지기반 임계값 제어기법에서 인식률과 구동시간을 멤버로 활용한 퍼지 제어기법은 데이터의 분류에서 높은 인식률을 유지하며 구동 시간을 절약하게 성능을 향상시키는 효과를 보여주었다.

향후 연구 방향은 우리가 제안한 군집화를 위한 퍼지기반 임계값 제어기법에서 인식률과 구동시간외의 군집 정보와 소속 데이터 및 소속 데이터의 수와 군집의 무게중심의 변화 등의 여러 분류 히스토리 정보를 활용하여 최적의 분류 임계값을 제어하는 다른 방법에 대한 연구를 진행 할 예정이다.

참고문헌

〔1〕 B. Ripley, "Pattern Recognition and Neural Networks," Cambridge, Mass.: Cambridge Univ. Press, 1996.

〔2〕 H.H. Bothe, "Fuzzy Logic," Springer, Berlin, Heidelberg, 1993

〔3〕 C.T. Lin and S.G. Lee, "Reinforcement structure / parameter learning for neural network based fuzzy logic systems," IEEE Trans. Fuzzy Systems, 46-63, 1994

〔4〕 S. Watson and S. Arkinstall, "The G-protein Linked Receptor Facts Book," Academic Press, Burlington, MA, 1994

〔5〕 D.T. Chalmers and D.P. Behan, "The Use of Constitutively Active GPCRs in Drug Discovery and Functional Genomics," Nature Reviews, Drug Discovery 1, p. 599-608, 2002

〔6〕 P. Cheng, Z. Ma, D. Cui, R.Geng, C. Chen, "Intelligent sequence adjusting algorithm based on general satisfaction function for air traffic arrival flow management," Computational Intelligence in Robotics and Automation, Proceedings. 533-537, July 2003

〔7〕 S. Gorinsky and H. Vin, "Extended Analysis of Binary Adjustment Algorithms," Technical Report TR2002-39, Department of Computer Sciences, The University of Texas at Austin, August 2002.

〔8〕 A. Devillez, P. Billaudel and G. Villermain Lecolier, "A fuzzy hybrid hierarchical clustering method with a new criterion able to find the optimal partition," Fuzzy Sets and Systems, p. 323-338, 2002

〔9〕 D. Tran, M. Wagner, "Generalised Fuzzy Hidden Markov Models for Speech Recognition," In Proceedings of the International Conference on Fuzzy Systems Calcutta, India. 2002

〔10〕 D. Tran and M. Wagner, "Fuzzy Clustering-Based Speaker Verification," Lecture Notes in Computer Science; Advances in Soft Computing - AFSS 2002, N.R. Pal, M. Sugeno(Eds.), 318-324, 2002

〔11〕 H. Takagi, N. Suzuki, T. Koda and Y. Kojima, "Neural networks designed on approximate reasoning architecture and there applications," IEEE Trans. Neural Networks, 752-760, 1992

〔12〕 C.T. Lin, Yeh. Chang-Moun, Hsu, Chun-Fei, "Fuzzy Neural Network Classification Design Using Support Vector Machine," IEEE International Symposium on Circuits and Systems, 724-727, 2004

〔13〕 G.A. Carpenter and S. Grossberg, "Adaptive resonance theory: Stable self-organization of neural recognition codes in response to arbitrary lists of input patterns," Proceedings of the 8th Conference of the Cognitive Science Society, Hillsdale, NJ: Erlbaum Associates, 45-62, 1988

〔14〕 J. S. R. Jang, C. T. Sun, E. Mizutani, "Neuro-Fuzzy and Soft Computing," Prentice-Hall Internationa l, 1997

〔15〕 C. W. de silva, "Intelligent Control," Fuzzy Logic Application. Boca Raton, FL: CRC, 1995

〔16〕 F. Horn, J. Weare, M. W. Beukers, Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F. and Vriend, G., "Gpcrdb: an information system for g protein-coupled receptors," Nucleic Acids Res., 26, 277-281, 1998

〔17〕 R. Karchin, K. Karplus, and D. Haussler. "Classifying g-protein coupled receptors with support vector machines," Bioinformatics 18:147-159, 2002

〔18〕 K..R. Sreekumar, et al, "Predicting GPCR-G-Protein coupling using hidden Markov models," Bioinformatics, 3490-3499, 2004

〔19〕 Jaakkola, and D. Haussler. "Exploiting generative models in discriminative classifiers," In Advances in Neural Information Processing Systems 11, Morgan Kauffmann, San mateo, Ca, 1998

〔20〕 H. Ying, and L. Yanda, "Classifying G-protein Coupled Receptors with Support Vector

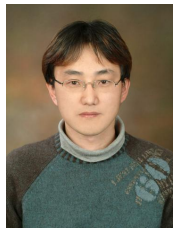
Machine," Advances in Neural Network(ISNN 2004), LNCS, p. 448 - 452, 2004

[21] J. Cheetham, F. Dehne, S. Pitre, A. Rau-Chaplin and P. J. Taillon, "Parallel clustal w for pc clusters," Proceedings of International Conference on Computational Science and Its Applications(ICCSA), vol. 2668, pp. 300 - 309, 2003

[22] B.P Zeigler, et al., "DEVS Framework for Modeling, Simulation, Analysis, and Design of Hybrid Systems," in Hybrid II, Lecture Notes in Computer Science, P. Antsaklis and A. Nerode, Editors. Springer-Verlag: Berlin. p. 529-551, 1996

[23] J. Nutaro and B. P. Zeigler, "On the Stability and Performance of Discrete Event Methods for Simulating Continuous Systems," Journal of Computational Physics, Vol 227, Issue 1, 10 November, 2007

저 자 소 개



조 규 철

2005 인하대학교 컴퓨터공학부 학사
2007 인하대학교 정보공학과 컴퓨터정보공학 석사
2007. 3~현재 인하대학교 정보공학과 컴퓨터정보공학 박사과정

관심분야 : 생명정보학, 소프트웨어 공학, 패턴인식, 최적화 알고리즘

마 용 범

2005 인하대학교 컴퓨터공학부 학사
2007 인하대학교 정보공학과 컴퓨터정보공학 석사
2007. 3~현재 인하대학교 정보공학과 컴퓨터정보공학 박사과정

관심분야 : 데이터 자동 분류, 그리드 컴퓨팅, 시뮬레이션

이 종 식

1993 인하대학교 전자공학과 학사
1995 인하대학교 전자공학과 석사
2001 미국 애리조나대 전기·컴퓨터공학과 박사
2001~2002 캘리포니아 주립대학교 전기·컴퓨터공학과 전임강사
2002~2003 클리블랜드 주립대학교 전기·컴퓨터공학과 조교수
2003~2006 인하대학교 컴퓨터공학부 조교수
2006~현재 인하대학교 컴퓨터공학부 부교수

관심분야: 시뮬레이션, 생명정보학, 패턴인식