Journal of The Korea Society of Computer and Information
Vol. 19, No. 8, August 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.8.011

클라우드 컴퓨팅 환경에서 무감독학습 방법과 퍼지이론을 이용한 결합형 데이터 분류기법

조 규 철*, 김 재 권**

Coupled data classification method using unsupervised learning and fuzzy logic in Cloud computing environment

Kyu-Cheol Cho*, Jae-Kwon Kim**

요 약

본 논문은 무감독학습을 통한 데이터 분류기법인 ART에서 퍼지이론을 이용한 결합형 데이터 분류 방법을 제안한다. 무감독학습기법 기반의 데이터 분류 기술은 분류기술의 향상의 장점이 있지만, 처리성능이 저하된다는 단점이 있다. 민첩성 있는 대용량데이터 처리와 분류인식률을 만족하는 최적의 임계값 결정기법이 필요하지만, 이는 불확실성이 많이 따르기 때문에 두 가지를 고려하여 상호보완 할 수 있는 처리기법이 필요하다. 제안하는 기법은 무감독학습을 하기 위해 퍼지매개변수와 퍼지 규칙을 설계하여 최적의 임계값을 도출한다. 제안하는 기법의 성능평가를 위해 클라우드 컴퓨팅환경에서 G 단백질 연결 수용체(G protein coupled receptor, GPCR)데이터를 이용하여 실험하였으며, 실험결과는 높은 인식률과 낮은 처리시간을 통해 결합형 데이터 분류에 효과적임을 입증하였다.

▸ Keywords : 클라우드 컴퓨팅, 결합형 데이터 분류, ART, 퍼지로직

Abstract

In This paper, we propose the unsupervised learning and fuzzy logic-based coupled data classification method base on ART. The unsupervised learning-based data classification helps improve the grouping technique, but decreases the processing efficiency. However, the data classification requires the decision technique to induce high success rate of data classification with optimal threshold. Therefore it is also necessary to solve the uncertainty of the threshold decision. The proposed method deduces the optimal threshold with the designing of fuzzy parameter and rules. In order to evaluate the proposed method, we design the simulation model with the GPCR(G

•제1저자 : 조규철 •교신저자 : 조규철
•투고일 : 2014. 7. 25, 심사일 : 2014. 8. 6, 게재확정일 : 2014. 8. 19.
* 한국전자통신연구원(Electronics and Telecommunications Research Institute)
** 인하대학교 컴퓨터정보공학과(School of Computer Science and Engineering, Inha University)

protein coupled receptor) data in cloud computing environment. Simulation results verify the efficiency of our method with the high recognition rate and low processing time.

▸ Keywords : Cloud Computing, Coupled data classification, ART, Fuzzy Logic

I. 서 론

빅 데이터(Big Data)의 이슈와 함께 대용량 데이터 분석을 위한 데이터 마이닝(Data Mining) 기반의 처리 방법이 주목받고 있다[1]. 클라우드 컴퓨팅(Cloud Computing)은 가상 자원을 이용하여 분산 환경에서의 대용량 데이터 처리가 가능한 기술이다. 클라우드 컴퓨팅을 활용한 데이터 마이닝은 대용량의 데이터를 여러 대의 가상 자원을 이용하여 빠르게 학습 할 수 있으며 의사결정이 가능하다[2][3][4]. 데이터 마이닝 기법에서 분류(Classification) 기술은 다양한 데이터에 대해 패턴정보를 추출하여 여러 산업분야에 적용되기 위해 지속적으로 연구가 진행되고 있다. 생물정보 데이터는 데이터의 정보화와 신규로 생성되는 방대한 데이터의 증가로 인해 대용량화되고 있으며 이를 처리하기 위한 데이터 마이닝 기법이 필요하다[5].

대용량의 데이터의 패턴을 분석하며 예측하기 위해서는 클라우드 가상화 환경을 기반으로 데이터를 처리 할 수 있는 기술이 요구된다. 하지만 클라우드 기술은 가상자원들을 동기화를 통해 메시지 및 메모리를 공유해야 되기 때문에 단일 작업을 수행하기 위한 협업 환경은 매우 어렵다. 또한, 데이터의 학습을 위해서는 각 데이터의 성격에 맞는 분산화된 학습 방법이 요구되고 있다. 무감독학습(Unsupervised Learning) 방법은 대용량데이터 처리를 위한 데이터 분류 시 미리 패턴의 종류를 결정하지 않고 주어진 데이터를 자동으로 학습하여 패턴을 결정하는 경우에 적합하다. 무감독학습 방법은 대표패턴 생성의 경계가 되는 임계값에 따라 생성되는 대표패턴의 수가 결정되며, 매개변수의 조정에 의해 두 개 이상의 패턴을 생성한다. 패턴의 수가 많아질 경우 가장 유사한 패턴을 찾을 수 있어 분류인식률이 향상된다는 장점이 있지만 데이터 처리 성능은 저하된다는 단점이 있다. 이러한 이유로 분류인식률에 영향을 최소화하면서 데이터 처리성능을 보장할 수 있는 최적의 임계값 결정은, 데이터 분류를 위해 중요한 이슈이다. 하지만 임계값의 결정은 불확실성(Uncertainty)이 많이 따르기 때문에 이를 처리하기 위한 기법이 필요하다.

본 논문은 ART(Adaptive Resonance Theory)[6]에서 퍼지이론을 통하여 최적의 임계값을 유도한다. 제안된 기법에 대한 성능을 실험하기 위해, 클라우드 컴퓨팅의 가상화 환경에서 대용량 생물정보 데이터인 G 단백질 연결 수용체 (G protein coupled receptor, GPCR) 데이터[7][8]를 이용하여 결합형 데이터 분류시스템 모델을 구현한다. 결합형 데이터 분류모델의 성능을 평가하기 위해 back-propagation, SVM과 제안하는 ART 데이터 분류기법과의 데이터 분류인식률 및 처리시간을 비교하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련연구에 대해 서술을 하며, 제 3장에서는 결합형 데이터 분류기법에 대해서 서술한다. 제 4장에서는 제안하는 기법에 대한 실험환경에 대해 서술을 하며, 제 5장에서는 실험 결과를 서술한다. 마지막으로 제 6장에서는 결론으로 마무리를 맺는다.

II. 관련 연구

1. 대용량 데이터 처리

대용량 데이터는 특정 도메인에 따라서 데이터가 비정형적인 데이터로 구성되어있으며, 기술적으로 일반적인 대용량 데이터 처리와 관련된 응용프로그램을 통한 데이터 처리는 분석과 해석이 어렵다. 또한, 비정형적인 대용량데이터 처리를 효과적으로 진행하기 위해 충분한 학습 데이터가 필요하지만, 학습데이터가 많아질 경우 학습 및 처리시간이 증가하게 된다. 이를 방지하기 위해 대부분의 대용량데이터 처리 응용프로그램들이 데이터를 군집화 하여 분산 환경에서 처리한다.

데이터 군집화는 기계학습 기법을 바꾸기보다는 연산능력과 저장 공간을 병렬적으로 처리하지만, 데이터 확장성에 대한 문제점을 가지고 있다. 이를 극복하기 위해 기계학습 기법을 개선하여 대용량데이터를 효율적으로 처리하는 방안으로 가상화 환경에서 분산처리 기반의 프레임워크와 분산 파일시스템에 대한 연구가 진행되어 왔다. NELL(Never-Ending Language Learning) 프로젝트는 연속적인(Continuous) 기계학습을 구축하는 것을 목표로 하여 비정형적인 대용량데

이터에서 구조화된 정보를 추출하고 있다. 또한, Ontology를 구축하여 자체 검증과정을 통해 구조화된 정보를 추출하고 학습하여, 네트워크상의 모든 콘텐츠에 대한 지식베이스를 구축하였다[9].

2. 데이터 처리를 위한 분류 기법

데이터 마이닝 기법을 활용하여 데이터 분류의 정확도를 향상시키는 연구는 꾸준히 진행 중이다. SVM(Support Vector Machine)[10]은 신경망 제어의 오류의 최소화를 목적으로 학습을 시키고, 패턴 인식 문제를 해결하기 위해 많이 사용되고 있다. 하지만 SVM은 두 개 이상의 데이터 셋을 훈련하는 경우, 1:N 다중 클래스를 구별하였다. 그러나 이러한 방법은 여러 요소를 고려한 데이터 분류가 불가능하고 자동학습이 불가능하다는 단점이 있다. ANN(Artificial Neural Network)은 사람의 신경망을 흉내내어 데이터를 학습하는 대표적인 알고리즘이다. ANN의 훈련과정 네트워크는 출력을 반복 수행하여 더욱더 정확한 결과를 유출해내는 이점이 있다. 그리나 ANN은 이상치가 데이터 내에 포함될 경우 입력자료를 평가할 수 없으며, 구조가 변경되는 경우 데이터 분류 및 학습이 불가능한 단점이 있다.

3. 임계값 제어 기법

무감독학습을 통한 데이터 분류는 패턴생성의 기준이 되는 임계값에 의해 분류인식률과 처리시간이 결정된다. 순차기반 임계값 제어기법[11]은 순차 기반의 결정범위를 제어하여 일정한 간격으로 임계값을 증가시켜 제어하는 기법이다. 임계값 조정 간격을 작게 하는 경우 분류인식률이 세밀하게 조절되어 요구하는 수준의 인식률 유도가 가능하지만, 임계값 갱신 시 도횟수가 증가할 경우 처리시간이 증가한다는 단점이 있다. 이진 기반 임계값 제어 기법[12]은 결정범위를 제어하는 방법이다. 임계값의 최소값과 최대값을 기준으로 조정 결정 범위를 정하고, 이전의 임계값에서 분류인식률이 낮을 경우 임계값 조정 결정 구간을 빼주고, 높을 경우 임계값 조정 결정 구간을 더해주어 분류인식률을 높일 수 있는 기법이다.

III. 결합형 데이터 분류 기법

1. 데이터 전처리와 결합형 데이터 분류 구현

결합형 데이터 분류를 이용한 GPCR 데이터의 분류 과정은 데이터 분류 과정과 데이터 재조정 과정으로 나눌 수 있다. 데이터 분류과정은 GPCR 데이터에 관한 대표패턴을 생성하고 대표패턴에 모든 데이터를 소속시킨 후 데이터를 분류하는 과정이다. 임계값 재조정 과정은 분류 인식률을 높이고 처리성능을 향상시키는 과정이다.

결합형 데이터 분류모델은 안정성 있는 분류인식률을 유도하고, 임계값 최적화를 통해 신뢰성 있는 데이터 분류를 진행할 수 있다. 훈련모듈은 학습을 통해 대표패턴을 생성하고 주어진 임계값에 따라 정의하여, 무감독학습을 통해 데이터 분류환경인 대표패턴들의 특징을 만들어 낸다. 테스팅 모듈은 이상치 감시기법을 통해 대표패턴의 의미가 없거나 특징이 없는 지식들을 여과한 후, 훈련모듈을 통해 생성된 새로운 데이터를 주입하여 유사한 패턴을 찾아내고 소속시켜 인식여부를 판단하게 된다. 입력데이터는 다양한 길이와 데이터형태가 다른 값들이기 때문에, 항상 안정성 있는 인식률을 제공해주지 못한다. 그러므로 재조정 모듈은 생성된 패턴에 대한 그룹화를 위해, 입력데이터를 읽어 데이터를 대표패턴과 비교가 가능하게 하는 재조정 모듈이 필요하다. 이때 데이터 재조정모듈은 대용량의 데이터를 가공하며 데이터를 준비하고 저장하는 단계가 있기 때문에 고비용이 필요하다.

2. 무감독학습 기반 데이터 분류 기법

대부분의 데이터 분류기법들은 2개의 대표 클래스를 기준으로 입력데이터의 특징을 정하고 데이터 분류를 진행한다. 하지만 본 논문에서 적용된 ART 기법은 무감독학습을 통해 여러 클래스에 대한 데이터 분류를 실시하여 기준 패턴을 정의하는 한계점을 극복한다. 그림 1은 GPCR 데이터를 분류하기 위한 ART기반의 무감독 학습 방법이다.

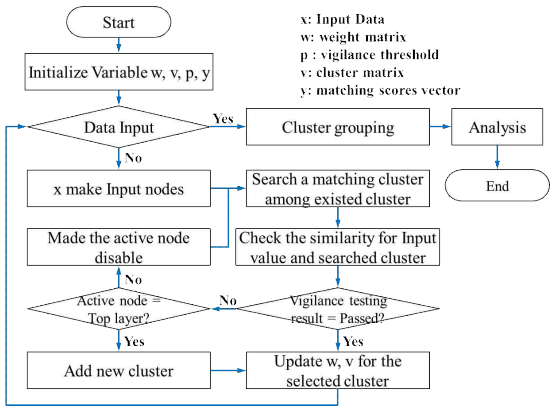


그림 1. ART 기반 분류 순서도
Fig. 1. ART based classification flowchart

ART는 기존에 학습되었던 것이 새로운 학습에 의해 지워지지 않도록 새로운 학습 지식을 자동적으로 전체 지식 베이스에 일관성 있게 통합하였다. 즉 적절하게 매치되는 새로운 학습 정보를 이용하여 기존에 배운 내용들을 정제하며, 새로운 인식 카테고리의 학습을 위해 새로운 패턴을 선택한다. 또한 기억용량을 넘어서는 과다한 새로운 입력에 대해서는 기존에 학습된 내용이 지워지는 것을 방지한다. 따라서 메모리 용량을 전부 소모할 때까지는 제한 없이 실시간으로 빠르고 안정되게 배울 수 있는 구조 이다.

3. 퍼지로직 기반 임계값 재조정

최적화된 분류 임계값은 처리시간의 최소값과 분류인식률에 대한 최대값을 보장하여 군집 정보를 조정하는 장점이 있는 것으로, 데이터의 형태와 학습량에 따라 유연하게 사용될 수 있다. 결합형 데이터 분류를 위한 퍼지기반 임계값 제어는 맘다니 모델을 이용하여 추론하고, 역퍼지화를 위해 무게 중심법(Center of Gravity)을 사용하였다.

본 논문에서 정의한 퍼지규칙과 멤버십 함수정의는 임의의 대용량데이터들에 대해서 데이터를 분류하는 시뮬레이션을 수행하여 임계값, 분류인식률, 데이터 처리시간을 측정하였다. 그리고 반복 수행한 시뮬레이션의 결과로써 설정된 임계값에 따른 분류인식률과 처리시간 결과값에 대한 분포를 기준으로 삼각형 멤버십 함수와 소속도값을 단위구간인 0~1사이의 값으로 정의하였다. 이는 애매한 규칙들을 경험과 시뮬레이션의 결과를 퍼지규칙에 통계적인 비율에 따른 분포된 값으로 적용함으로써, 최적 임계값 유도를 위해 신뢰성을 제공하고 임계값 설정에 따라 예상되는 결과에 대한 오차를 줄일 수 있기 때문이다. 퍼지논리를 이용한 임계값 제어를 위한 퍼지기법의 입력 매개변수는 데이터 분류에서의 분류인식률(X), 데이터 처리시간(Y)과 임계값(Z)이다. 표 1은 본 논문에서 사용된 각 퍼지집합의 소속도 함수를 보여준다. 표에서 소속도 함수정의는 입력 값에 대한 멤버십 함수가 있고 멤버십 함수에 대한 구간을 설명영역에서 정의하고 있다. 설명에서 맴버십 함수의 구간으로서 삼각 퍼지함수를 이용한다. 퍼지함수는 구간의 입력값과 소속도값을 표현한 것으로, L은 삼각형의 왼쪽을 나타내고, M은 삼각형의 중간부분, R은 삼각형의 오른쪽을 나타낸다.

그림 2는 데이터 분류를 위한 임계값을 갱신하기 위해 60가지의 퍼지규칙을 정하여, 분류인식률과 데이터 처리시간에 관한 입력 멤버십함수와 임계값을 정하기 위한 퍼지 규칙을 나타낸 것이다.

표 1. 퍼지 매개변수
Table 1. Fuzzy parameter value

파라미터	맴버십 함수	맴버십 함수 소속값
분류 인식률 (X)	LOW	M:80%, R:90%
	NORMAL	L:80%, M:91.5%, R:95%
	HIGH	L:93%, M:95% R:99%
	VERY HIGH	L:96%, M:100%
처리시간 (Y)	FAST	M:0s R:6s
	NORMAL	L:4s, M:8s, R:16s
	SLOW	L:10s, M:20s, R:∞
임계값 (Z)	VERY LOW	L:0, M:1, R:1.4
	LOW	L:1.2, M:1.4, R:1.8
	NORMAL	L:1.4, M:1.6, R:1.8
	HIGH	L:1.6, M:1.8, R:2.0
	VERY HIGH	L:1.8, M:2.0, R:∞

임계값(Z)		VERY LOW	LOW	NORMAL	HIGH	VERY HIGH
처리시간(Y)	분류 인식률(X)					
SLOW	LOW	Rule1 VERY LOW	Rule2 VERY LOW	Rule3 LOW	Rule4 NORMAL	Rule5 LOW
	NORMAL	Rule6 VERY LOW	Rule7 LOW	Rule8 NORMAL	Rule9 NORMAL	Rule10 NORMAL
	HIGH	Rule11 VERY LOW	Rule12 LOW	Rule13 NORMAL	Rule14 NORMAL	Rule15 NORMAL
	VERY HIGH	Rule16 LOW	Rule17 LOW	Rule18 NORMAL	Rule19 HIGH	Rule20 HIGH
NORMAL	LOW	Rule21 VERY LOW	Rule22 LOW	Rule23 LOW	Rule24 LOW	Rule25 NORMAL
	NORMAL	Rule26 VERY LOW	Rule27 LOW	Rule28 NORMAL	Rule29 NORMAL	Rule30 NORMAL
	HIGH	Rule31 VERY LOW	Rule32 LOW	Rule33 NORMAL	Rule34 HIGH	Rule35 VERY HIGH
	VERY HIGH	Rule36 LOW	Rule37 LOW	Rule38 NORMAL	Rule39 HIGH	Rule40 VERY HIGH
FAST	LOW	Rule41 VERY LOW	Rule42 LOW	Rule43 NORMAL	Rule44 HIGH	Rule45 VERY HIGH
	NORMAL	Rule46 VERY LOW	Rule47 LOW	Rule48 NORMAL	Rule49 HIGH	Rule50 VERY HIGH
	HIGH	Rule51 VERY LOW	Rule52 NORMAL	Rule53 HIGH	Rule54 VERY HIGH	Rule55 VERY HIGH
	VERY HIGH	Rule56 LOW	Rule57 NORMAL	Rule58 HIGH	Rule59 VERY HIGH	Rule60 VERY HIGH

그림 2. 임계값 제어를 위한 퍼지 규칙
Fig. 2. Fuzzy Rule for Threshold control

식 1은 분류인식률과 처리시간 값을 퍼지이론에 적용하여 다음 경계 임계값을 책정하기 위한 수식이다. n은 퍼지 입력에 적합한 퍼지규칙의 수이고, zi는 퍼지규칙 전건부의 연산결과인 소속도 값을 의미한다. 그리고 Tmax는 퍼지규칙의 출력 퍼지집합이 가질 수 있는 최대 유일 값을 의미한다.

$$T = \sum_{i=1}^{n} (T_{\max i} \times Z_i) / \sum_{i=1}^{n} Z_i \quad (1)$$

예를 들어 분류인식률이 91%, 데이터 처리시간이 11, 임계값이 1.75라고 가정하자. 분류인식률(X)의 값이 91%이면 X값에 대한 퍼지규칙은 'NORMAL'이고 소속도값은 0.83을 가리킨다. 처리시간(Y)의 값이 11이면 Y값에 대한 퍼지규칙은 'NORMAL, 'SLOW'이고, 이에 대한 소속도값은 각각 0.53과 0.07이다. 또한 임계값(Z)의 값이1.75이면 Z값에

대한 퍼지규칙은 'NORMAL, 'HIGH'이고, 이에 대한 소속도 값은 각각 0.23과 0.82가 된다. 이 경우 퍼지 입력 값들은 그림 3의 퍼지추론 규칙 내에 Rule8, 9, 28, 29에 적합하다. Rule8, 9, 28, 29에 해당되는 z8, z9z28, z29의 소속도 값은 각각 0.07, 0.23, 0.53, 0.82이고 출력된 임계값은 1.9636이다. 결합형 데이터 분류의 임계값은 임의의 입력패턴과 저장된 패턴을 통해 불일치 허용도가 결정된다. 임계값을 범위 내에서 자동적으로 최적의 값을 변환하도록 유도함으로써, 분류인식률과 데이터 처리시간을 조절하며 최적의 데이터 분류환경을 조절할 수 있게 된다.

IV. 실험 환경 및 구현

1. 실험 환경

클라우드 컴퓨팅환경에서 대용량데이터 처리를 하기 위해 DEVS[13] W/S(Discrete Event System Specification Web Service)환경을 구축하여 실험한다. 실험을 위한 데이터 분류는 각 컴포넌트화 되어 있는 각 모듈과 서비스를 통합형태로 결합하여 모델링 하였다. DEVS W/S는 분산 시뮬레이션, 모델간 상호운영, 복잡한 시뮬레이션이 가능하고, 각 모델간의 시간동기화와 사용자 인터페이스는 웹환경을 지원한다. 또한 복잡한 다량의 정보를 송수신하고 양방향 정보 전달이 가능하다는 이점이 있다. 시간서버 중심으로 STSI (Simulation Time Synchronization Infrastructure)가 구성되어 있다. 데이터 분류 시뮬레이션 시간을 웹서비스 실행단위에 참여한 모든 참여컴포넌트에게 전송하여 진행한다.

본 논문에서는 제안하는 기법을 입증하기 위해서 3가지의 실험을 진행한다. 첫 번째 실험은 데이터 분류 인식률로써, 해당하는 기법이 분류 인식률의 성능을 측정한다. 두 번째 실험은 데이터 분류 시간으로, 분류 시간이 GPCR 데이터를 사용하였을 경우에 분류 속도를 위해 측정한다. 마지막으로 임계값 횟수에 따른 구동시간 변화에 대해 측정한다. 이는 임계값이 낮아짐과 높아짐에 따라 분류 시간과 인식에 대해 얼마나 합당한지 측정한다.

2. 3-tier 기반 데이터 처리 설계

그림 3과 같이 3-tier 기반의 데이터 처리는 임계 값을 통해 인식률을 보장하고 안정적인 성능을 보장하는 데이터 분류 환경이다. 분산컴퓨팅환경에서 데이터 분류를 위한 자원들을 하나의 컴포넌트로 구성하여, 대용량데이터 처리를 위해 확장된 데이터 저장능력과 분류체계를 갖추고 있다. 3-tier 기반 데이터 처리는 에이전트와 데이터 훈련 및 임계값을 조정하는 기능들을 각 자원에 할당하여 자원의 역할을 배분하여 수행한다. 본 논문에서는 GPCR 데이터를 이용하여 데이터 분류기법에 대한 분류인식률을 측정하였다. GPCR 데이터는 현재 알려져 있는 2,000개의 데이터에 대해 최상위 클래스인 A, B, C, D 및 decoy 클래스로 분류가 되어 있다[7][8]. 실험을 위한 데이터는 분류가 되어있는 모든 데이터를 섞어 각 분류기법을 통하여 다시 데이터를 클래스로 분류하였다. 그리고 각 분류기법의 결과와 GPCR 데이터 정의에서 클래스 소속이 맞게 배치가 되어 있는 경우 데이터 분류 성공으로, 다른 클래스에 배치되어 있는 경우 데이터 분류 실패로 간주한다.

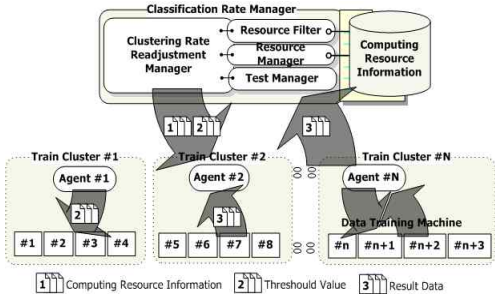


그림 3. 3-tier 기반 데이터 처리
Fig. 3. 3-tier based data processing

V. 실험 결과

1. 데이터 분류 인식률

SVM, back-propagation, ART분류, ART 이상치 감시기법(ARTOD)에 대해서 데이터 분류 실험을 실시하였고, 수식 2을 기준으로 인식률의 결과를 측정하여 그 성능을 비교하였다. (분류인식률(CSR: Classification Success Rate), 분류성공(CS), 분류실패(CF))

$$CSR = \frac{Count(CS)}{Count(CS) + Count(CF)} \times 100 \quad (2)$$

GPCR 데이터에 대해서는 A클래스그룹과 A클래스를 제외한 나머지 그룹을 나머지 클래스그룹(The others class)으로 정의하여 데이터 분류를 진행하였다. 검증된 분류기법인 back-propagation, SVM과 ART분류 기법과의 성능을 비

교한다. 전처리를 통해 변형된 744개의 GPCR 데이터에 대해서 임의로 100개의 셋으로 추출하였다. 100개의 데이터 셋은 훈련 데이터 셋과 테스트 데이터 셋을 구분하여 실험을 진행하였는데, 각각의 A 클래스그룹과 나머지그룹의 훈련 데이터는 572개이고, 테스트 데이터는 62개를 사용하였다. 그 결과 100개의 데이터 셋을 통해 A 클래스그룹과 나머지 클래스그룹에 대한 테스트 결과로 744건의 분류 결과를 얻었다. 그림 4는 이에 대한 결과로써 4가지의 분류 기법에 대한 GPCR 분류인식률을 표현하였다.

A 클래스그룹에 대한 결과는 744건의 테스트 중, SVM 기법은 742건, back-propagation기법은 741건, ART 및 ARTOD 기법은 742건의 분류성공을 보여, 네 개의 분류기법 모두 99.5% 이상의 분류인식률을 제공하였다. 4가지 분류기법에 대한 A 클래스그룹의 데이터분류는, 거의 모든 데이터에서 분류 성공하였음을 알 수 있다. 나머지 클래스그룹에 대한 결과는 SVM기법이 720건, back-propagation기법은 719건, ART분류와 ARTOD 분류기법은 719건의 분류를 성공하였고, 이는 모든 분류기법이 96.5% 이상의 분류성공을 통해 A 클래스그룹과 나머지 클래스그룹에 대한 분류성능이 탁월함을 알 수 있었다. 결합형 데이터 분류기법의 데이터 분류성능이 나머지 분류 기법의 성능과 비슷한 분류성능을 제공하고 있고, GPCR 데이터의 분류에 유용하게 사용할 수 있다는 것을 알 수 있다.

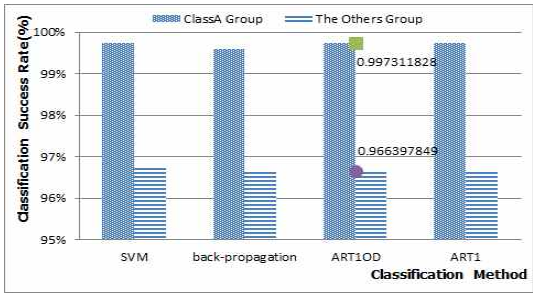


그림 4. 분류 인식률
Fig. 4. Classification success rate

2. 데이터 분류 시간

데이터 분류 시간에 대한 실험은 데이터 분류 인식률의 결과를 통해 안정성이 검증된 데이터 분류기법으로, SVM, back-propagation, ART 및 ARTOD기법에 대해서 데이터 분류 시 평균 처리시간을 비교하였다. 평균 분류시간이 짧다는 것은, 대용량데이터 분류에서 짧은 데이터 처리시간으로 더 많은 데이터 분류를 진행할 수 있다는 것이고, 데이터 분류를 위해 활용되는 컴퓨팅자원을 절약할 수 있다는 것이다. 이는 자원의 사용률을 향상시키고 서비스 시간에 대한 예측오차를 줄일 수 있다는 기대효과가 있다고 볼 수 있다.

데이터 분류의 데이터 처리시간은 GPCR 데이터의 학습시간과 테스트시간을 합친 것이다. 데이터 분류시간은 GPCR 데이터 양, 단일 데이터 크기, 임계값에 따라 생성되는 대표패턴의 수에 따라 달라진다. 그림 5는 GPCR 데이터에 대한 데이터 분류를 진행하여 SVM, back-propagation, ART 및 ARTOD기법의 반복 횟수와 대표패턴 비율을 조절함으로써, 데이터 분류를 위한 처리시간을 측정한 결과이다.

대표패턴 생성비율이 증가함에 따라 데이터 처리시간도 증가하는데, 데이터의 수와 생성되는 대표패턴의 수가 많아질수록 모든 분류기법은 처리시간을 절약함을 알 수 있다. 특히 ARTOD기법은 back-propagation기법보다 43%, SVM기법보다 53%, ART기법보다 12%의 데이터처리시간을 절약하고 있다. 따라서, 결합형 데이터 분류기법이 다른 기법보다 향상된 성능을 제공함을 알 수 있고, 대표성이 떨어지는 패턴 여과를 통해 처리성능을 향상시키고 있음을 알 수 있다. 결합형 데이터 분류기법은 데이터 분류를 위한 빠른 성능을 제공하고, 참여자원에 대한 사용 효율을 향상시킨다는 것을 알 수 있다. 또한 실험의 결과는 할당된 작업에 대하여 예상 처리시간을 최소화할 수 있는 기대효과가 있음을 입증하고 있다.

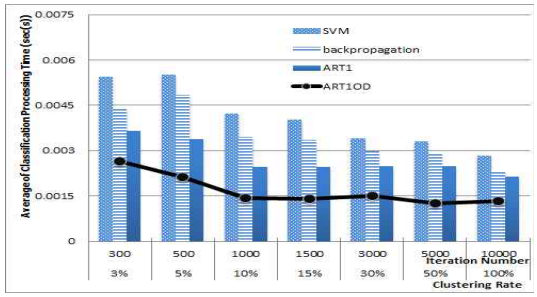


그림 5. 데이터 분류 시간
Fig. 5. Classification processing time

3. 임계값 변화 횟수에 따른 구동 시간 변화

무감독학습을 통한 데이터 분류는 패턴생성의 기준이 되는 경계기준인 임계값에 의해 분류인식률과 처리시간이 결정된다. 임계값이 낮은 경우에는 생성되는 패턴의 수가 적기 때문에 패턴 비교시간을 절약하여 빠른 처리시간을 제공하지만, 임계값이 높아지는 경우 생성되는 많은 비교 패턴으로 인해 처리시간이 증가하게 된다. 따라서 데이터 분류를 위해 최적 임계값을 유도한다는 것은 분류인식률과 처리시간을 결정하기 위한 중요한 요소라고

할 수 있다. 본 실험은 1개의 데이터 셋에 대해 95%~96% 분류 인식률을 제공하는 최적의 임계값을 유도하는 시간을 측정하였다. 퍼지기반 임계값 제어(Fuzzy-based Threshold)기법을 통한 최적의 임계값 유도성능을 평가하기 위해, 순차기반 임계값 제어(Sequence-based Threshold)기법과 이진기반 임계값 제어(Binary-based Threshold)기법을 비교하였고, 그림 6과 7은 최적의 분류성공률을 유도하는 시뮬레이션간의 임계값 변화와 분류성공률의 변화를 도식화한 것이고 그림 8은 임계값 갱신 횟수에 따른 구동 시간의 변화를 측정하여 도식화한 것이다.

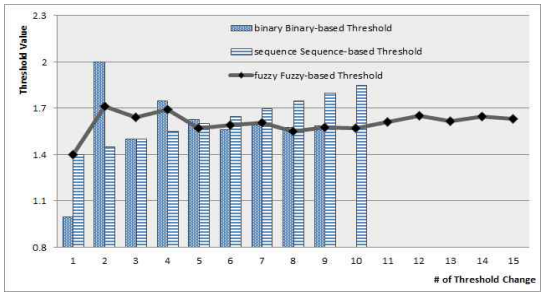


그림 6. 최적 분류를 위한 임계값 변화
Fig. 7. Threshold change for the optimal data classification

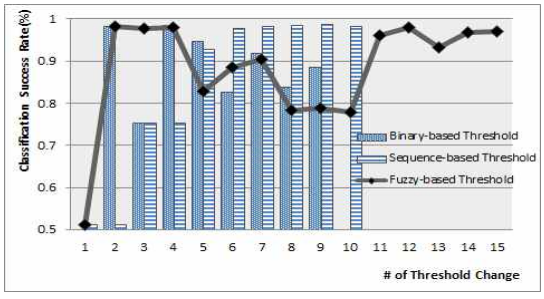


그림 7. 임계값에 따른 분류 성공률
Fig. 8. Classification success rate according to threshold change

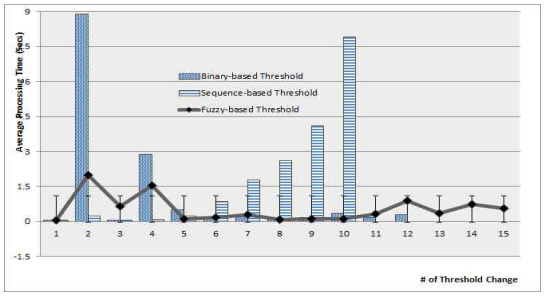


그림 8. 임계값 변화에 따른 평균 처리시간 변화
Fig. 9. Threshold change follow the average processing time change

순차기반 임계값 제어기법은 초기에 분류 임계값이 낮기 때문에 생성되는 군집의 수가 적어 낮은 데이터 처리시간을 기록하였다. 하지만 분류 임계값이 높아져 생성되는 패턴의 수도 많아지고 구동시간이 길어져서 비효율적이라 할 수 있다. 이진 기반 임계값 제어기법은 초기의 이진 임계값 갱신 폭이 높기 때문에 많은 처리시간이 요구되지만, 4번의 임계값 갱신 이후에는 약 13초의 수렴된 값을 통해 안정된 처리시간을 제공하고 있다. 또한 퍼지기반 임계값 제어기법은 초기부터 낮은 증가율을 제시하며, 다른 두 제어기법보다 낮은 처리시간을 기록하였다. 특히 4번의 임계값 갱신 이후에는 약 5초에 수렴되는 값을 통해 처리시간을 절약하기 위해 효율적인 기법인 것을 입증하고 있다.

IV. 결 론

네트워크의 발전과 정보화 기술의 발전으로 데이터는 대용량화 되어 가고 있으며 대용량 데이터를 효율적으로 관리하고 처리할 수 있는 기법이 필요하다. 대용량의 데이터를 분석하기 위해서는 각 도메인에 맞는 분류 기법이 필요하며, 데이터 마이닝 기술을 이용한다. 본 논문에서는 분류 인식률의 최소화를 위해서 ART 기반의 퍼지를 활용한 임계값 조정방법을 제안하였다. 제안하는 기법은 결합형 데이터 분류기법으로서 분류 인식률을 높이며 처리시간을 효율적으로 줄일 수 있었으며, GPCR 데이터를 통해 그 성능을 입증하였다. 제안하는 결합형 데이터 분류 기법은 GPCR데이터 인식뿐만 아니라 대용량 데이터가 처리 가능한 도구로써 다양한 분야에 응용 가능하다. 이 기법을 이용하여 무감독학습기법이 요구되는 대용량 데이터 처리환경에 유용할 것으로 기대된다.

향후 연구로는 결합형 데이터 분류기법에서 데이터 섹션단위의 데이터 분류 방안을 도출하고, 단위데이터가 큰 데이터 분류에 대한 처리시간과 메모리 문제에 대한 문제해결을 위한 연구를 지속할 것이다.

참고문헌

[1] Y. Zhang, D. Sow, D. Turaga and M. Schaar, "A Fast Online Learning Algorithm for Distributed Mining of BigData," ACM SIGMETRICS performance Evaluation Review, Vol. 41, Issue 4, pp. 90-93, March 2014.

[2] Xue, Liangfei, Dongfeng Yuan, and Mingyan

Jiang, "Web Data Mining Based on Cloud Computing," Proceedings of the 2012 International Conference on Cybernetics and Informatics. Springer New York, 2014.

[3] Cho D.K. and Park S.C., "Development and Implementation of Monitoring System for Management of Virtual Resource Based on Cloud Computing," Journal of The Korea Society of Computer and Information, Vol. 18, No. 2, pp. 41-47, 2013

[4] Kang I.S., Kim T.H. and Lee H.C., "Data processing techniques applying data mining based on enterprise cloud computing," Journal of the Korea society of computer and information, Vol. 16, No. 8, pp. 1-10, 2011.

[5] Kim J.K., Lee J.S., Park D.K., Lim Y.S., Lee Y.H. and Jung E.Y., "Adaptive mining prediction model for content recommendation to coronary heart disease patients", Cluster Computing, 2013. DOI: 10.1007/s10586-013-0308-1

[6] Stephen Grossberg, "Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world," Neural Networks, Vol. 37, pp. 1-47, 2013.

[7] F. Horn, J. Weare, M. W. Beukers, S. Hörsch, A. Bairoch, W. Chen, Ø. Edvardsen, F. Campagne and G. Vriend, "GPCRDB: An Information System for G Protein-Coupled Receptors," Nucleic Acids Res, Vol. 26, Issue 1, pp. 275-279, 1998.

[8] D. T. Chalmers and D. P. Behan, "The use of Constitutively Active GPCRs in Drug Discovery and Functional Genomics," Nature Reviews, Drug Discovery, Vol 1, No. 8, pp. 599-608, 2002.

[9] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr., and T. M. Mitchell, "Toward an Architecture for Never-Ending Language Learning," Proceeding of the Conference on Artificial Intelligence AAAI Press, Vol. 5, pp. 1306~1313, 2010.

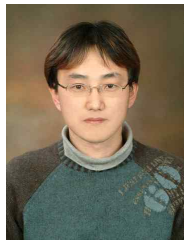
[10] Z. Qi, Y. Tian and Y. Shi, "Robust twin support vector machine for pattern classification," Pattern Recognition, Vol. 46, Issue 1, pp. 305-316, 2013.

[11] P. Cheng, Z. Ma, D. Cui, R. Geng and C. Chen, "Intelligent Sequence Adjusting Algorithm Based on General Satisfaction Function for Air Traffic Arrival Flow Management," Proceeding of the Computational Intelligence in Robotics and Automation, pp. 533-537, 2003.

[12] 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, 2002.

[13] B. P. Zeigler, H. S. Song, T. G. Kim and H. Praehofer, "DEVS Framework for Modeling, Simulation, Analysis, and Design of Hybrid Systems in Hybrid," Lecture Notes in Computer Science, Vol. 999, pp.529-551, 1995.

저 자 소 개



조 규 철

2005: 인하대학교
컴퓨터공학과 공학사.
2007: 인하대학교
컴퓨터정보공학과 공학석사.
2013: 인하대학교
컴퓨터정보공학과 공학박사
현 재: 한국전자통신연구원 근무 중
관심분야: 클라우드 컴퓨팅, 시뮬레이션
Email : kccho@etri.re.kr

김 재 권

2011: 가천의과학대학교
정보처리과 공학사.
2013: 인하대학교
컴퓨터정보공학과 공학석사.
현 재: 인하대학교
컴퓨터정보공학과 박사과정
관심분야: 클라우드 컴퓨팅, 인공지능
Email : jaekwonkorea@naver.com