韓國 컴퓨터情報學會 論文誌
第13卷 第1號, 2008. 1.

2008-13-1-1-5

지역적 컨셉트 적응형 IOLIN시스템을 사용한 데이터 스트림의 분류

김 재 우*, 송 재 원**, 이 주 홍***

Data Streams classification using Local Concept-adapted IOLIN System

Jae-Woo Kim *, Jae-Won Song **, Ju-Hong Lee ***

요 약

데이터 스트림은 시간이 경과함에 따라서 데이터의 패턴이 변화하는 특성이 있다. 데이터 스트림에 내재되어 있는 이러한 특성(컨셉트 변화)은 분류 모델의 예측 성능을 감소시킨다. CVFDT와 IOLIN은 점진적인 분류모델의 갱신을 통해 컨셉트 변화를 해결하고자 하였다. 그러나 이러한 방법들은 작은 패턴의 변화가 전체 분류 결과에 영향을 주는 지역적 컨셉트 변화를 식별하지 못함으로써 모델을 재구축하는 단점이 있다. 본 논문은 컨셉트 변화 발생 시 지역적 컨셉트 변화를 찾음으로써 시스템의 예측 성능을 향상시키는 적응형 IOLIN을 제안한다. 실험 결과는 제안 기법인 적응형 IOLIN기법이 IOLIN기법에 비해 정확률에서 약 2.8%, CVFDT기법보다 약 11.2%정도 우수하였다.

Abstract

Data stream has the tendency to change in patterns over time. Also known as concept drift, such problem can reduce the predictive performance of a classification model. CVFDT and IOLIN tried to solve the problem of a concept drift through incremental classification model updates. The local changes in patterns, however, was revealed to be unable to resolve the problems of local concept drift that occurs by influencing on total classification results. In this paper, we propose adapted IOLIN system that improves system's predictive performance by detecting the local concept drift. The experimental result shows that adaptive IOLIN, the proposed method, is about 2.8% in accuracy better than IOLIN and about 11.2% in accuracy better than CVFDT.

▸ Keyword : 실시간 데이터 마이닝(Real-Time Data Mining), 온라인 학습(Online Learning), 컨셉트 변화(Concept Drift), 정보 퍼지 네트워크(Info-Fuzzy Network).

• 제1저자 : 김재우 교신저자 : 이주홍
• 접수일 : 2007. 12. 14, 심사일 : 2008. 1.15, 심사완료일 : 2008. 1.21.
* 인하대학교 컴퓨터・정보공학과 석사과정 **인하대학교 컴퓨터・정보공학과 박사과정
***인하대학교 컴퓨터・정보공학과 부교수
※ 이 논문은 인하대학교의 지원에 의하여 연구되었음.

www.kci.go.kr

Ⅰ. 서 론

데이터 마이닝에서 분류문제는 데이터의 특성을 분석하는 주요 연구 분야이다. 대부분의 전통적인 데이터 마이닝 분류 기법들은 데이터의 패턴이 변하지 않는 정적 데이터의 집합을 처리 대상으로 하였다. 그러나 최근 하드웨어 기술과 인터넷, 센서 네트워크, 위성통신등과 같은 다양한 형태의 통신 기술의 발달과 함께 스트림 형태의 데이터가 발생하였다. 전통적인 데이터 집합과는 달리 데이터 스트림은 일시적 순서(temporally order), 빠른 변화(fast changing), 잠재적으로 무한한 크기 등과 같은 특징들을 가지기 때문에 기존의 마이닝 분류 기법을 적용시키기에는 많은 장애가 있었다[1]. 따라서 데이터 스트림을 분류하기 위해 최적화된 기법들이 필요하다. 최근 데이터 스트림 분류에 관한 다양한 연구들이 진행되고 있다[1, 2].

데이터 스트림을 분류 시 발생하는 가장 큰 장애요인은 시간이 경과함에 따라서 이전에 구축되었던 분류 모델의 성능이 급격히 낮아지는 컨셉트 변화(concept drift)이다[3]. Aggarwal은 이를 데이터 스트림 진화(data stream evolution)라고 표현하였다[4]. 컨셉트 변화는 분류 모델의 예측 성능을 감소시키며, 시간 복잡도와 요구되는 메모리의 양을 증가시킨다. 때문에 스트림 데이터의 분류 모델들은 시간이 지남에 따라 입력 데이터와 패턴이 다른 현상을 보인다. 분류모델과 입력 데이터의 컨셉트의 차이는 효율적인 데이터 스트림 분류의 장애요인으로 지적된다. 따라서 현재의 분류 모델과 입력 데이터의 컨셉트의 차이를 해결하기 위해서 최근의 입력 데이터를 효과적으로 반영하는 모델로 재구축해야한다. 컨셉트 변화는 데이터 패턴의 변화 범위에 따라서 지역적 컨셉트 변화(local concept drift)를 구분할 수 있다[2]. 지역적 컨셉트 변화는 입력 데이터의 대부분의 속성들이 유효함에도 불구하고 작은 패턴의 변화가 전체 데이터의 분류에 영향을 주어 정확률을 현저하게 감소시키는 현상을 말한다.

컨셉트 변화를 고려한 기존 데이터 스트림 기법들로 컨셉트가 적용된 매우 빠른 의사결정나무(CVFDT ; Concept-adapting Very Fast Decision Tree)[5], 온라인 정보 네트워크(OLIN ; OnLine Information Network)[6], 점진적 온라인 정보 네트워크(IOLIN ; Incremental OnLine Information Network)[7] 들이 있다. 위의 기법들에서 OLIN과 IOLIN은 컨셉트 변화를 해결하기 위하여 새로운 모델을 구축한다. 그러나 지역적 컨셉트 변화를 고려하지 못하였기 때문에 실행시간의 지연과 정확률의 감소가 발생한다.

본 논문은 컨셉트 변화 시 지역적 컨셉트 변화를 찾음으로써 분류 모델을 효율적으로 구축하는 적응형 IOLIN시스템을 제안한다. 제안된 적응형 IOLIN은 지역적 컨셉트 변화 문제를 해결함으로써 모델 구축의 빈발도를 낮추고 실행시간을 단축시켰다. 적응형 IOLIN의 기본적인 알고리즘은 컨셉트 변화가 발생하지 않으면 기존 모델을 갱신하고, 컨셉트 변화가 발생하면 지역적 컨셉트 변화인지 구분한다. 만약 지역적 컨셉트 변화라면 모델을 갱신하고, 그렇지 않다면 새로운 모델을 구축한다.

제안된 기법의 타당성을 검토하기 위하여 실험 데이터를 통해 CVFDT와 IOLIN에 대한 정확률과 실행시간에 대하여 비교 실험 하고 그 결과를 제시한다. 제안된 적응형 IOLIN기법은 IOLIN보다 빠른 실행시간을 보였다. 또한 정확률에 있어서도 CVFDT, IOLIN과 비교하여 좀 더 나은 성능을 보여주었다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 데이터 스트림에서의 분류 방법과 컨셉트 변화를 적용한 분류방법을 소개한다. 3장에서는 본 논문에서 제안하는 지역적 컨셉트 적응형 IOLIN 알고리즘을 기술한다. 4장에서는 컨셉트 적응형 IOLIN 알고리즘과 기존의 시스템과의 성능을 비교하였다. 마지막 5장에서는 결론 및 향후 연구방향에 대해서 기술한다.

Ⅱ. 관련 연구

2.1 VFDT와 CVFDT

트리기반의 VFDT[8]는 가장 널리 알려진 데이터 스트림 분류기법이다. VFDT의 장점은 트리기반의 구조를 사용함으로써 데이터의 빠른 처리속도를 갖는다. 그러나 컨셉트 변화에 의해 갑작스럽게 정확률이 떨어지는 문제를 해결할 수 없었고, 이에 대한 연구가 진행되었다[5].

Hulten[5]은 스트림 데이터의 컨셉트 변화를 다루기 위해서 VFDT를 확장한 CVFDT(Concept adapting Very Fast Decision Tree)를 제안하였다. CVFDT의 갱신 방법은 새로운 데이터의 패턴을 반영한 부분트리(sub-tree)를 새롭게 만들고 기존의 부분트리와 주기적인 비교를 통해 정확률을 평가한다. 기존의 부분트리보다 새로 구축한 부분트리가 최근 데이터 패턴을 더 잘 분류할 때, 기존의 부분트리를 새로 구축한 부분트리로 교체한다.

www.kci.go.kr

2.2 정보 퍼지 네트워크(IFN)

정보 퍼지 네트워크(IFN; Info-Fuzzy Network)는 정보 이론(Information Theory)을 배경으로 속성들 간의 다중 계층 네트워크를 구축하는 데이터 분류 모델이다. IFN은 정보 이론(데이터의 속성을 분리하는 기준)을 사용한 방법들 중의 하나로서 Last & Maimom[9]에 의해 개발되었다.

IFN은 입력속성들과 출력 속성 사이의 상호정보(MI; Mutual Information)를 계산하기 위하여 다중 계층 네트워크(multi-layered network)를 구축한다. 입력속성들과 출력속성 사이의 조건적인 MI값을 구하여 가장 높은 값을 갖는 속성이 첫 번째 층이 된다. 이후 MI값이 0이 될 때까지 조건적인 MI값이 높은 순서대로 층을 구성한다. MI값이 0이라면 노드는 더 이상 분리되지 않고 어느 한 클래스로 분류 된다. 따라서 IFN모델은 조건적인 MI값이 0이 되거나 더 이상 계층을 만들 수 있는 속성이 없을 때까지 모델을 구축한다. 이 시스템은 네트워크로서 의사결정트리(decision tree)와 같이 목표 속성(target attribute)을 예측하는데 사용되지만, 모든 단말노드(leaf node)들은 목표 층(target layer)의 모든 노드들과 네트워크로 연결되는 차이점을 갖고 있다. 그림 1은 3계층 구조의 IFN을 보여준다. 단말노드 2, 1.2, 1.1.1, 1.1.2는 목표 노드 0, 1과 연결됨으로써 네트워크 구조를 이룬다. 이는 트리 구조와 가장 큰 차이점이다.

IFN은 분류를 결정하는데 필요한 입력 속성들의 수가 다른 알고리즘들에 비해 상대적으로 적어 빠른 데이터의 처리가 가능하다. 그 이유는 분류를 결정짓는 속성들을 잘 선별함으로써 입력 속성의 수를 줄이기 때문이다.

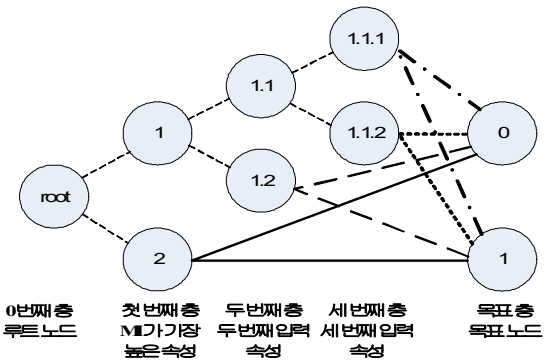

그림 1. 3계층 구조의 정보 퍼지 네트워크
Fig. 1 Info-fuzzy network three layered structure

2.3 온라인 정보 네트워크(OLIN)

Last[6]는 IFN을 온라인 상에서 구현한 OLIN 시스템을 제안하였다. IFN을 기반으로 한 OLIN 시스템은 컨셉트 변화에 따라 윈도우의 크기를 동적으로 조절하여 연속적인 스트림 데이터를 처리한다. 그림 2는 OLIN 시스템을 도식화한 것이다. 만약 데이터 스트림 상에서 컨셉트 변화가 발생한다면, 윈도우의 크기를 줄임으로써 컨셉트 변화에 따른 오류율을 줄이고 새로운 IFN을 구축한다. 정확률이 높으면 컨셉트 변화가 발생하지 않았다고 판단하여 윈도우의 크기를 확대하고, 새로운 IFN을 구축한다. 이러한 동적인 윈도우 크기의 변화는 고정된 크기를 가진 윈도우보다 더 높은 정확률을 보여 주었다[6]. 그러나 OLIN 시스템은 새로운 데이터가 입력될 때마다 컨셉트 변화에 상관없이 새로운 네트워크를 구축하기 때문에 많은 구축비용을 요구한다.

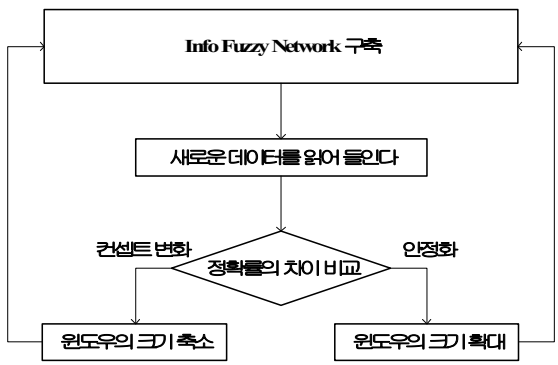

그림 2. 온라인 정보 네트워크 시스템
Fig. 2 OnLine Information Network system

2.4 점진적 온라인 정보 네트워크(IOLIN)

Cohen[7]은 실행시간이 느린 OLIN의 성능을 향상시키기 위하여 IOLIN을 제안하였다. IOLIN은 OLIN과 달리 최근 데이터를 이전 모델에 반영함으로써 모델구축에 따른 비용과 시간을 절약하였다. IOLIN은 OLIN에서의 실험을 바탕으로 개선되었다. OLIN에서 이전 모델과 현재 모델의 차이점을 비교한 결과, 컨셉트 변화가 발생하지 않았을 때, 이전에 구축되었던 모델과 현재 구축된 모델의 80%가 유사함을 보였다. 나머지 20%에서 모델의 마지막 층이 현저하게 다른 점을 발견하였다. 이러한 실험 결과를 반영하여, IOLIN은 컨셉트 변화가 발생하지 않았을 때, 현재 모델의 마지막 층만을 갱신한다. 따라서 IOLIN은 OLIN보다 데이터를 분류하는

www.kci.go.kr

정확률은 약간 낮지만 훨씬 빠른 실행속도를 보여준다.

그림 3은 IOLIN 시스템을 도식화한 것이다. 데이터 스트림 상에서 컨셉트 변화가 발생한다면, 윈도우의 크기를 줄이고 새로운 IFN을 구축한다. 컨셉트 변화가 발생하지 않는다면, 현재 IFN을 갱신하고 윈도우의 크기를 확대한 후에 새로운 IFN을 구축하지 않고 다음 데이터를 읽어 들인다. 하지만 IOLIN은 지역적 컨셉트 변화를 처리하는 절차가 없기 때문에, 컨셉트 변화 시 매번 새로운 네트워크를 구축함으로써 효율성이 떨어지는 문제점을 가지고 있다.

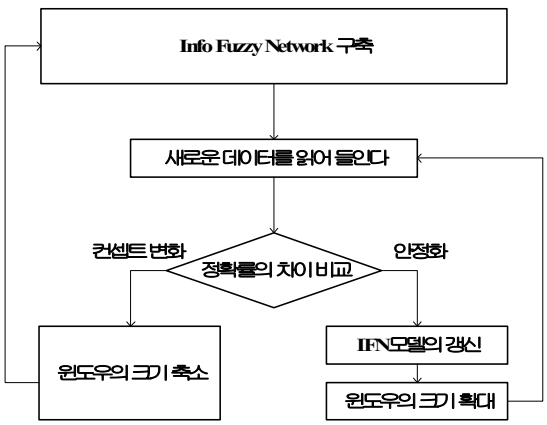

그림 3. 점진적 온라인 정보 네트워크 시스템
Fig. 3 Incremental OnLine Information Network system

III. 지역적 컨셉트 적응형 IOLIN

3.1 지역적 컨셉트 적응형 IOLIN 알고리즘

본 절에서는 IOLIN의 지역적 컨셉트 변화 시 성능저하 문제를 해결하기 위하여 적응형 IOLIN을 제안한다. 제안된 적응형 IOLIN은 컨셉트 변화 시 지역적 컨셉트 변화를 식별한 후 입력 데이터에 맞춰 모델을 갱신한다. 이러한 지역적 컨셉트 변화를 구분하는 처리절차를 시스템에 반영함으로써 분류정확률을 높이고, 빈번한 모델 재구축 비용을 줄임으로써 실행시간을 감소시킨다. 다음은 제안된 지역적 컨셉트 변화에 따른 적응형 IOLIN알고리즘이다.

표 1의 적응형 IOLIN의 알고리즘은 결과 값인 갱신된 IFN을 얻기 위해 입력 값으로 현재 IFN과 윈도우의 크기(W)를 입력 받는다. 표 1 알고리즘의 4번째 줄은 패턴의 변화에 따른 컨셉트 변화를 측정하기위해 Maxdiff를 계산한다. Maxdiff는 이항 분포를 사용하여 훈련 집합과 실험 집합의 정확률간의 차를 이용하여 구한다.

표 1. 적응형 IOLIN
Table 1. Adapted IOLIN

	적응형 IOLIN(*IFN*, *W*)
1:	스트림 데이터 입력
2:	E_{tr} ← 훈련 집합의 오류율
3:	E_{val} ← 실험 집합의 오류율
4:	Max_{diff} ← E_{tr} 와 E_{val} 사이의 최대 오차율 계산
5:	**if** $E_{tr} - E_{val} \geq Max_{diff}$ **then**
6:	윈도우의 크기를 확장
7:	*IFN* 모델에 입력 데이터를 반영하여 갱신
8:	**else**
9:	윈도우의 크기를 축소
10:	C_{mi}← 입력 데이터의 MI가 가장 높은 속성
11:	F_{mi}← *IFN*모델의 첫 번째 층의 조건적인 MI
12:	**if** $F_{mi} \leq C_{mi}$ **then**
13:	새로운 *IFN*모델을 구축
14:	**else**
15:	지역적 컨셉트 변화 처리 실행
16:	**end if**
17:	**end if**
18:	**return**
19:	갱신된 *IFN*모델

Maxdiff를 구하는 식은 (1),(2)와 같다:

$$Var_{diff} = \frac{E_{tr}(1-E_{tr})}{W} + \frac{E_{val}(1-E_{val})}{Add_{Count}} \quad \cdots\cdots\cdots (1)$$

$$Max_{diff} = z_{0.99}\sqrt{Var_{diff}} = 2.326\sqrt{Var_{diff}} \quad \cdots\cdots (2)$$

식(1)에서 Addcount는 W에 가감되는 윈도우의 크기이며, Etr와 Eval은 훈련 집합의 오류율(error rate)과 현재 입력된 데이터의 오류율이다.

컨셉트 변화는 두 오류율 Etr와 Eval간의 차를 Maxdiff와 비교하여 구분한다. 만약 Etr-Eval의 값이 Maxdiff보다 작다면, 컨셉트 변화가 발생하지 않았다고 판단하고 기존의 모델에 입력 데이터를 반영하여 IFN을 갱신한다[7]. 하지만 그 차이가 Maxdiff보다 크다면 컨셉트 변화가 발생했다고 판단하고, Fmi와 Cmi의 비교를 통해 지역적 컨셉트 변화를 구분한다.

표 1의 12번째 줄은 지역적 컨셉트 변화를 구분하는 방법이다. Cmi는 입력 데이터의 조건적인 MI가 가장 높은 속성

www.kci.go.kr

이고, Fmi는 IFN모델의 첫 번째 층의 조건적인 MI 값이다. 만약 Fmi 보다 Cmi의 값이 더 크다면 새로운 IFN을 구축한다. 그렇지 않다면, 지역적 컨셉트 변화가 발생하였다고 판단하여 지역적 컨셉트 변화처리 절차를 실행한다. IFN은 구조상 가장 높은 MI 값을 갖는 속성이 첫 번째 층으로 구축된다[6, 9]. 또한 이 첫 번째 층은 스트림 데이터를 분류하는데 있어서 40~50%의 영향을 준다[6]. 그렇기 때문에 첫 번째 층을 이루는 속성의 MI 값이 입력 데이터의 가장 높은 MI 값보다 작다면, 새로운IFN모델을 구축하는 것이다. 조건적인 MI[6]를 구하는 식은 (3)과 같다:

$$MI(A_{i'}, A_i / z) = \sum_{j=0}^{M_i-1}\sum_{j'=0}^{M_{i'}-1} P(V_{ij}; V_{i'j'}; z) \cdot \log\frac{P(V_i^{ij}_{j'} / z)}{P(V_{i'j'}/z) \cdot P(V_{ij}/z)} \quad \cdots (3)$$

표 2는 식(3)에서 MI를 구성하는 기본 요소들이다.

표 2. 조건적인 MI의 기본 요소
Table 2. Basic elements of Conditional Mutual Information

기호	의미
A_i	목표 속성
$A_{i'}$	주어진 노드 z에서의 후보 입력 속성
V_{ij}	속성 A_i의 값 j
$P(V_{ij} / z)$	주어진 노드 z에서 평가된 V_{ij}의 조건 확률
$P(V_{i'j'}^{ij} / z)$	주어진 노드 z에서 평가된 V_{ij}와 $V_{i'j'}$의 조건 확률
$P(V_{ij} ; V_{i'j'} ; z)$	주어진 노드 z에서 목표 속성 i의 값 j와 노드 z와 후보 입력 속성 i'의 값 j'의 평가된 결합 확률

3.2 지역적 컨셉트 처리 절차

지역적 컨셉트 처리 절차는 이전 데이터의 속성과 입력 데이터의 변화 시 속성의 조건적인 MI값을 비교한다. 이를 통해 입력데이터의 지역적 패턴 변화를 식별함으로써 IFN모델을 부분적으로 갱신한다. 아래 표 3은 지역적 컨셉트 변화 시 처리절차이다.

표 3. 지역적 컨셉트 변화 시 절차
Table 3. procedure when Local concept drift detecting

지역적 컨셉트 변화 처리(*IFN*)
for i = 2 **to** i = 마지막 층
F_m← *IFN*모델의 현재 층의 조건적인 MI
C_m← 입력 데이터의 조건적인 MI가 가장 높은 속성
if $F_m(i)$*95%≤$C_m(i)$ **then**
i 번째 층의 $F_m(i)$를 $C_m(i)$의 속성으로 교체
분리 검증 조사를 실행
return
else
현재 층을 유지하고 다음 층으로 이동
end if
if i 가 마지막 층이라면 **then**
분리 검증 조사를 실행
end if

표 3의 지역적 컨셉트 변화 시 절차는 두 번째 층부터 마지막 층까지 IFN 모델의 현재 층의 조건적인 MI(Fmi)와 입력 데이터의 조건적인 MI (Cmi)를 비교한다. 만약 Cmi가 더 높은 값을 갖는다면, Cmi의 속성을 현재 층으로 만들고, 현재 층 이하의 모든 층을 삭제하고 분리 검증 조사를 실행한다. 만약 Fmi값이 더 크다면 현재 층을 유지하고 다음 층으로 이동한다. 표 3의 4번째 줄의 95%는 통계의 유의확률(significant probability)로서 유의수준은 0.05로 설정한다. 이러한 MI값의 비교를 통해 지역적 컨셉트 변화가 발생한 층을 알 수 있다.

분리 검증 조사[7]는 각 노드의 분리기준이 실험 집합으로부터 계산한 조건적인 MI에 실제로 영향을 받는지를 검증한다. 표 4는 분리 검증 조사 절차이다.

표 4. 분리 검증 조사 절차
Table 4. New Split Validation Process

분리 검증 조사(*IFN, W*)
for j = 1 **to** j = 은닉층 i 에서의 노드의 수
if 노드 j 가 분리된다면 **then**
j와 목표 속성의 평가된 조건적인 MI를 계산
j의 가능도비(likelihood ratio)를 계산
가능도비 통계의 자유도를 계산
if j의 가능도비 확률 > 노드의 분리 기준점 **then**
노드 j를 분리 표시
else
분리된 노드를 제거하고 j를 단말노드로 만듦
end if
end if

www.kci.go.kr

표 4의 4번째 줄은 현재 노드가 자식 노드를 갖는지 가능도비 검증 (likelihood-ratio test)[6]을 계산한다. E*(z)는 노드 z와 연관된 레코드들의 개수이다:

$$G^2(A_{i'};A_i/z) = 2 \bullet (\ln 2) \bullet E^*(z) \bullet MI(A_{i'};A_i/z) \quad \cdots (4)$$

표 4의 5번째 줄은 가능도비 통계의 자유도[6]를 계산한다. α는 네트워크 노드를 나누기 위한 최솟값으로 사용자 정의가 가능하며 실험에서는 0.1%를 기본 값으로 정의하였다:

$$G^2(A_{i'};A_i/z) \geq \chi^2_\alpha((NI_{i'}(z)-1) \bullet (NT_i(z)-1)) \quad \cdots\cdots (5)$$

IV. 실험 및 결과

지역적 컨셉트 적응형 IOLIN의 성능평가 방법은 IOLIN과 CVFDT알고리즘을 비교 실험하여 정확률과 실행시간을 측정하였다. 실험 환경은 Pentium 4 D 3.2G 프로세서와 1GB 램, Windows XP Professional 운영체제와 150GB 하드를 사용하였다. 개발언어는 자바를 사용하였다.

데이터 집합은 대신증권[13]에서 얻은 코스피 200(KOSPI200; Korea Stock Price Index 200)에 포함된 100개 회사들의 3년 동안의 일일 주식가격을 가지고 평가한다. D1과 D2는 정형 데이터 집합이고, D3, D4, D5는 컨셉트 변화를 갖는 비정형 데이터 집합이다.

4.1 분류 정확률 비교 실험

첫 번째 실험 평가에서는 데이터 스트림 분류의 정확률을 측정 하였다. 그림 4는 정확률의 결과이다. CVFDT보다 IOLIN과 적응형 IOLIN이 높은 정확률을 보였다. 그림 4의 D1과 D2 데이터 집합에서 IOLIN과 적응형 IOLIN의 그래프가 같은 이유는 컨셉트 변화가 발생하지 않았을 때, 두 알고리즘은 같은 절차를 실행하기 때문이다. 그러나 지역적 컨셉트 변화를 갖는 데이터가 입력된다면, IOLIN은 새로운 모델을 구축한다. 이러한 방법은 이전 데이터의 정보를 모두 버림으로써 입력 데이터에 최적화시키는 방법이다. 하지만 이러한 방법은 컨셉트 변화가 심한 데이터에서는 과체적화(overfitting) 현상이 발생되어 모델의 성능을 떨어뜨린다. 적응형 IOLIN은 지역적 컨셉트 변화의 상황에서 단지 갱신만을 수행하며, 이는 이전 데이터의 정보를 유지할 수 있게 해준다. 적응형 IOLIN은 IOLIN보다 2.8% 높은 평균 정확률을 보여주었다. 또한 CVFDT보다 평균 11.2% 정도 우수하였다.

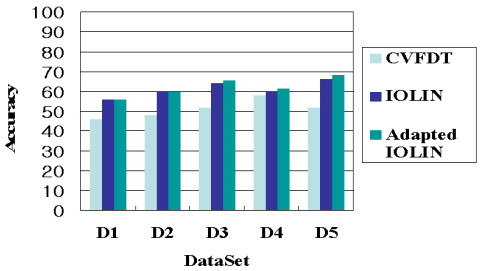

그림 4. 분류 정확률 측정 결과 비교
Fig. 4 Comparison of Classification Accuracy

4.2 실행시간 비교 실험

두 번째 실험 평가에서는 데이터 스트림 분류의 실행시간을 비교하였다. 그림 5는 실행시간의 실험결과이다. CVFDT 알고리즘이 가장 빠른 실행시간을 보여주었는데, 그 이유는 데이터가 원패스(one-pass) 알고리즘을 사용하는 단순한 처리방법 때문이다. 지역적 컨셉트 변화가 발생했을 때, IOLIN이 컨셉트에 적응형인 IOLIN보다 실행시간이 느린 이유는 새로운 모델을 구축하는데 드는 비용 때문이다. 적응형 IOLIN은 IOLIN보다 모델 구축의 빈발도를 10%이상 낮출 수 있었고, 이는 IOLIN보다 제안한 시스템이 더욱 빠른 실행시간을 가짐을 보여주었다. 제안된 시스템은 속도에서 IOLIN보다 평균 9.8% 향상된 속도를 나타내었다.

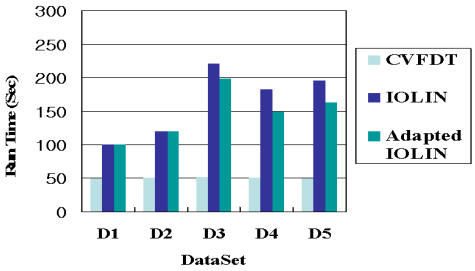

그림 5. 실행시간 측정 실험 결과 비교
Fig. 5 Comparison of Run Time

www.kci.go.kr

V. 결론

데이터 스트림 분류에 있어서 가장 근본적인 문제는 시간이 경과에 따라 입력 데이터의 패턴이 변하는 컨셉트 변화이다. 지역적 컨셉트 변화는 컨셉트 변화 시 작은 패턴의 변화가 전체 데이터의 분류에 영향을 주는 현상이다. 우리는 컨셉트 변화에서 지역적 컨셉트 변화를 판별하여 분류의 정확률과 실행시간을 향상시키는 지역적 컨셉트 적응형 IOLIN를 제안한다.

지역적 컨셉트 적응형 IOLIN은 컨셉트 변화 상황에서 처리속도에 있어서 평균적으로 IOLIN보다 좀 더 빠른 처리속도를 보였다. 또한 정확률에 있어서도 CVFDT와 IOLIN에 비해 높은 성능을 보여주었다. 이는 제안된 시스템이 지역적 컨셉트 변화를 고려함으로써 모델 구축 시 소비되는 시간을 줄였기 때문이다. 뿐만 아니라 변화된 부분 패턴을 모델에 반영함으로써 분류 정확률을 높였다.

참고문헌

〔1〕 C. Aggarwal, Data Streams: Models and Algorithms, 354Page, Springer, 2007.

〔2〕 A. Tsymbal, The problem of concept drift: definitions and related work, Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College Dublin, Ireland, 2004.

〔3〕 G. Widmer and M. Kubat, Learning in the Presence of Concept Drift and Hidden Contexts, Machine Learning, Vol. 23, No. 1, pp. 69-101, 1996.

〔4〕 C. Aggarwal, A Framework for Diagnosing Changes in Evolving Data Streams. Proceedings of the ACM SIGKDD Conference, 2003.

〔5〕 G. Hulten, L, Spencer, and P. Domingos, "Mining Time-Changing Data Streams", Proc. of KDD 2001, pp. 97-106, ACM Press, 2001.

〔6〕 M. Last, "Online Classification of Nonstationary Data Streams", Intelligent Data Analysis, Vol. 6, No. 2, pp. 129-147 2002.

〔7〕 L. Cohen, M. Last, G. Avrahami, "Incremental Info-Fuzzy Algorithm for Real Time Data Mining of Non-Stationary Data Streams", TDM Workshop, Brighton UK, 2004.

〔8〕 P. Domingos and G. Hulten. "Mining high-speed data streams" In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 71-80, Boston, MA, 2000. ACM Press.

〔9〕 O. Maimon and M. Last, Knowledge Discovery and Data Mining - The Info-Fuzzy Network (IFN) Methodology, Kluwer Academic Publishers, December 2000.

〔10〕 L. Cohen, G. Avrahami, M. Last, A. Kandel, and O. Kipersztok, "Incremental Classification of Nonstationary Data Streams", Proceedings of the Second International Workshop on Knowledge Discovery in Data Streams, pp. 117-124, October 7, 2005, Porto, Portugal.

〔11〕 R. Klinkenberg, Learning drifting concepts: example selection vs. example weighting, Intelligent Data Analysis, Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, 8 (3), 2004.

〔12〕 H. Wang, W. Fan, Yu P.S., Han J., Mining concept-drifting data streams using ensemble classifiers, Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining KDD-2003, ACM Press, 2003, 226-235.

〔13〕 daisin securities web site at http://www.daishin.co.kr/

www.kci.go.kr

저 자 소 개

김 재 우

2006년 2월 : 명지대학교 컴퓨터 소프트웨어학과 학사

2006년 ~ 현재 : 인하대학교 대학원 석사과정

관심분야 : 컨셉트 변화, 데이터마이닝, 데이터스트림

송 재 원

2005년 2월 : 성공회대학교 전산정보학과 학사

2007년 2월 : 인하대학교 대학원 석사

2007년 ~ 현재 : 인하대학교 대학원 박사과정

관심분야 : 데이터베이스, 데이터마이닝, 정보검색

이 주 홍

2001년 2월 : 한국 과학 기술원 컴퓨터 공학 박사

2002년 ~ 현재 : 인하대학교 컴퓨터 공학부 부교수

관심분야 : 데이터마이닝, 데이터베이스, 정보검색, 신경망, 기계학습

www.kci.go.kr