실시간 비즈니스 프로세스 모니터링 방법론을 위한 확장 KNN 대체 기반 LOF 예측 알고리즘

Extended KNN Imputation Based LOF Prediction Algorithm for Real-time Business Process Monitoring Method

강복영(Bokyoung Kang)*, 김동수(Dongsoo Kim)**, 강석호(Suk-Ho Kang)***

초 록

본 논문에서는 KNN 대체와 LOF 알고리즘의 결합 모델을 확장하여 실시간 비즈니스 프로세스 모니터링을 위한 비정상 종료 예측 방법론을 제안하였다. 기존의 룰 기반 모니터링 방법론은 실시간 프로세스 진행 정도에 따른 비관측 정보에 기인하여 조기 경보 및 실시간 대응이 힘들다는 한계점을 안고 있다. 이를 해결하기 위하여 비관측 정보에 대한 가정 및 진행 중인 프로세스의 향후 경로 예측을 통해 종료 시점에서 예상되는 LOF를 추정하기 위한 알고리즘을 제안하였다. 이 알고리즘을 적용하여 실시간 비즈니스 프로세스 모니터링 과정에서 각 관측 시점마다 종료 시점에서의 결과를 예측함으로써, 전 시점에 걸친 추세를 살펴 종료 패턴을 예측할 수 있다. 이를 통해 비즈니스 프로세스의 실시간 진척에 대한 정보를 가시화함으로써 기회 및 위협에 사전에 대응할 수 있게 하여 프로세스 관리 수준의 향상을 기대할 수 있을 것으로 예상된다.

ABSTRACT

In this paper, we propose a novel approach to fault prediction for real-time business process monitoring method using extended KNN imputation based LOF prediction. Existing rule-based approaches to process monitoring has some limitations like late alarm for fault occurrence or no indicators about real-time progress, since there exist unobserved attributes according to the monitoring phase during process executions. To improve these limitations, we propose an algorithm for LOF prediction by adopting the imputation method to assume unobserved attributes. LOF of ongoing instance is calculated by assuming next probable progresses after the monitoring phase, which is conducted during entire monitoring phases so that we can predict the abnormal termination of the ongoing ins-

본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2010-0020943).

* 서울대학교 산업공학과 박사과정

** 교신저자, 숭실대학교 산업정보시스템공학과 부교수

*** 서울대학교 산업공학과 교수

2010년 11월 16일 접수, 2010년 11월 22일 심사완료 후 2010년 11월 23일 게재확정.

tance. By visualizing the real-time progress in terms of the probability on abnormal termination, we can provide more proactive operations to opportunities or risks during the real-time monitoring.

키워드 : LOF(Local Outlier Factor), 대체 기법, 실시간, 프로세스 모니터링, 이상치 감지 및 예측
LOF, Imputation, Real-time, Process Monitoring, Outlier Detection and Prediction

1. 서 론

비즈니스 프로세스 모니터링 시스템은 기업의 목적 달성을 위해 비즈니스 프로세스를 수행하는 동안 내, 외부에서 발생하는 다양한 사건(event)이나 관련 속성 지표(attribute)와 같은 정보들에 대한 실시간 접근성을 제공하기 위한 정보 시스템이다[2]. 비즈니스 프로세스는 비즈니스 목표를 달성하기 위한 태스크의 집합으로 구성되어 있다[11]. 프로세스를 실행하면 인스턴스가 발생하고, 이 인스턴스는 프로세스 및 태스크의 수행 과정에서 기록된 관련 속성들로 관측된다. 이러한 속성들은 각 태스크의 수행과 관련된 시간, 입력 및 출력 자료, 리소스, 각종 오류 상황 등의 정보를 통칭한다[7]. 모니터링 시스템은 과거에 수행된 프로세스 로그의 분석을 통해, 속성들의 패턴과 프로세스의 상태 간의 관계를 도출하여 룰로 정의하고, 이는 추후에 실행되는 프로세스 인스턴스에서 발생하는 정보의 관측을 통해 인스턴스의 상태를 규명할 수 있게 한다[21].

이러한 목적을 위하여 데이터마이닝을 적용한 룰 기반 프로세스 모니터링 방법론에 관한 연구들이 널리 활용된다. 과거에 수행된 프로세스 로그를 데이터 마이닝 기법을 이용하여 분석함으로써, 인스턴스의 수행 과정에서 발생하는 프로세스 관련 데이터와 종료 결과 간의 상관관계를 “If(조건부)-Then(결과 및 그에 따른 대응 방안의 수행)” 형태의 룰로 추출한다[7, 8]. 룰 기반 모니터링의 과정은 룰의 조건부에 대한 감지(detecting)와 그에 해당하는 상태의 규명 및 그에 따른 대응(predefined operation)의 제공으로 이루어진다.

본 연구에서의 프로세스 모니터링의 주요 목적은 빈번하게 수행된 정상적인 프로세스의 수행상태와 상이한 희귀 패턴(infrequent pattern)을 감지하는 것, 즉 이상치 감지(anomaly detection)로 한정한다. 전체 과거 사례들을 구성하는 빈번하고 중요한 데이터의 패턴과 다른 상태로 수행된 소수의 또는 새로운 인스턴스를 찾아내어 비정상적인 프로세스의 종료를 감지함으로써 그로 인한 손해에 대한 대응책을 마련할 수 있다. 이러한 이상치 탐지는 여러 분야에서 중요성이 인식되고 있으며, 특히, 공정의 진행 상태 모니터링[4], 신용카드 부정사용 등의 금융사기 탐지[22], 보험사기 방지[20], 응급 진료에서의 이상 상태 진단[16], 영상 관측을 통한 희귀 패턴 감지[15] 등에서 활용되고 있다.

최근, 이상치 감지를 위한 알고리즘으로 높은 성능을 보이는 Local outlier factor(LOF) algorithm[1]을 적용한 연구가 활발히 제안되

고 있다. LOF 알고리즘은 밀도기반의 이상치 탐지 방식으로 다차원 공간상에서의 이상치는 그렇지 않은 개체들이 갖는 주변 밀도에 비해 주변 밀도가 극히 낮다는 성질을 이용해 이상치를 탐지한다. LOF 알고리즘은 이러한 주변 밀도에 근거하기 때문에, 일반적인 거리 기반의 기존 알고리즘에 비해 전체 데이터 개체에 대해 이상 정도를 말해주는 하나의 단일 지표값인 LOF를 계산할 수 있고, 전체 이상치가 아니라 부분 이상치에 대해서도 탐지가 가능하기 때문에 다른 알고리즘보다 일반적으로 높은 성능을 보인다[13].

하지만, LOF를 포함한 룰 기반 모니터링 방법론들은 실시간 비즈니스 프로세스 모니터링에 적용될 경우 한계점이 발생한다. 기존의 제안 방법론들은 모든 속성 정보가 확보되어야만 룰 기반의 상태 규명이 가능해진다. 따라서, 인스턴스의 모든 정보가 프로세스의 종단에서 순간적으로 또는 빠르게 수집되는 경우, 인스턴스 단위의 룰 기반 상태 규명이 원활히 이루어질 수 있다. 하지만, 프로세스의 대형화 및 복잡화에 의해 단일 인스턴스의 수행에 소요되는 시간이 길어질 수 있으며, 이때는 데이터의 수집이 프로세스의 진행 과정에 걸쳐 단계적으로 수집된다[10]. 따라서, 단일 인스턴스의 진행 과정을 실시간으로 모니터링 할 경우, 인스턴스가 종료될 때까지 기다려야만 하는 상황이 발생한다[7, 8]. 그러므로 인스턴스가 비정상적인 종료가 예정된 상태로 진행되더라도 모니터링 시스템은 이미 비정상적 종료가 발행한 뒤, 그에 대한 보상을 위한 대응적 조정(reactive correction)만을 수행할 수 있다[12]. 이러한 한계점으로부터, 실시간으로 진행 중인 인스턴스의 추후에 발생 가능한 다양한 결과들을 예측함으로써 주도적인 예방(proactive prevention)이 가능한 실시간 모니터링 방법론 개발에 필요함을 알 수 있다[14].

이를 해결하기 위해 본 논문에서는 LOF를 이용하여 비정상적인 프로세스의 종료를 주도적으로 예측하기 위한 실시간 프로세스 모니터링 방법론을 제안하였다. 진행 중인 인스턴스를 실시간으로 모니터링 하는 동안, 각 모니터링 시점에는 해당 시점까지 수집된 속성 정보만을 알고 있어, LOF 값의 계산이 불가능하고 인스턴스의 상태 규명 또한 이루어질 수 없다. 우리는 대체(Imputation) 기법을 도입하여 관측 시점 이후의 비획득 정보(unknown attributes), 즉 진행 중인 인스턴스의 추후 수행 과정에서 관측될 것이라 예상되는 속성정보를 가정하였다. 이로써, 현재 상태 및 그에 기반한 향후 경로의 예측을 통해 LOF 계산이 가능하도록 하였다. 이를 이용하여 매 관측 시점마다 인스턴스의 종료 시점에서 예상되는 LOF의 분포를 추정할 수 있으며 이는 인스턴스가 진행 되는 동안 전 관측 시점에 걸쳐 수행된다. 따라서, 경과 과정에 따른 추세를 살펴 비정상적인 종료를 미리 예측함으로써 보다 주도적인 실시간 프로세스 모니터링이 가능하다.

2. 배경 연구

본 연구에서는 룰 기반 모니터링을 위해 활용되는 이상치 감지 알고리즘들 가운데, 비주도학습(unsupervised) 기법들 중 일반적으로 가장 높은 성능을 보이는 LOF 알고리즘

을 이용하고 있다. LOF 알고리즘의 기본 아이디어는 각각의 전체 데이터 개체에 대해 개별적인 개체 마다 이상치 정도를 나타내는 측정치를 계산하는 것이다. 이러한 측정치를 LOF라 한다. 다른 개체들에 비해 가장 가까운 주변 개체들과의 밀도가 낮게 측정되는 개체는 높은 LOF 값을 가지며 강한 이상치임을 암시하게 되고, 다른 개체들에 비해 가장 가까운 주변 개체들과의 밀도에 큰 차이가 없으면 정상 개체로 판단한다. LOF 알고리즘은 다음의 절차로 이루어진다[1].

1. 모든 데이터 개체 q에 대해 k-distance(q)를 q와 q의 k-th nearest neighbor 사이의 거리로 정의 하고 계산한다.
2. 모든 데이터 개체 p와 q사이의 reach-$dist_k$(q, p)를 아래의 식으로 구한다. 단, d(q, p)는 p와 q사이의 유클리드 거리, k는 이상치로 판단시키지 않게 하고 싶은 최소 군집의 크기를 의미한다.
   reach-$dist_k$(q, p)=max{d(q, p)$_k$-distance(p)}
3. 데이터 개체 q의 local reachability density(lrd) 계산한다. 단, Nk(q)는 q의 k nearest neighbor 집합이다. 개체 q의 lrd는 위 식에서와 같이 개체 q의 k nearest neighbor 집합에 속하는 개체들의 평균적인 reachability distance의 역수이다.

$$lrd_k(q) = 1/\left(\frac{\sum_{p\in N_k(q)} reach-dist_k(q,p)}{|N_k(q)|}\right)$$

4. 데이터 개체 q의 Local Outlier Factor (LOF)의 계산한다.

$$LOF_k(q) = \frac{\sum_{p\in N_k(q)} \frac{lrd_k(p)}{lrd_k(q)}}{|N_k(q)|} \quad (1)$$

데이터 개체의 LOF 값은 자기 자신의 밀도와 주변 개체들의 밀도와 비교되어 주변 개체들과의 밀도가 비슷한, 다시 말해 군집 밖에 위치한 개체에 대해서는 그 값이 식 (3)에 따라 1로 근접하게 되고, 주변 개체들과의 밀도가 자신의 이웃 개체들의 밀도에 차이가 많이 나는 경우에는 그 정도에 비례하여 1을 상회하는 값을 갖게 된다. 따라서, 인스턴스의 수행 과정에서 관측된 각 속성 값들로 이루어진 벡터로부터 LOF를 계산함으로써, 해당 인스턴스가 얼마나 비정상적인지의 정도를 수치화할 수 있다.

LOF를 비롯한 이상치 감지 알고리즘들을 이용한 룰 기반 모니터링 접근법들은 실시간 프로세스 모니터링에 활용될 경우, 다음과 같은 한계점이 발생한다. 기존의 접근법에서는 모든 프로세스 관련 속성 정보가 완전히 확보된 상황에서 If-Then 룰에 근거한 상태의 규명 또는 분류(identification or classification)가 이루어진다. LOF 또한 모든 각 샘플의 모든 데이터가 확보되어야만 알고리즘 수행이 가능하다[17]. 하지만, 실시간 프로세스 모니터링에서는 프로세스 인스턴스가 진행되는 과정에서 프로세스를 구성하는 태스크들의 수행 과정에 따라 각 태스크로부터 관련 속성 정보가 순차적으로 발생하고 이 값들이 정보 시스템에 의해 수집된다[10]. 기존의 기법들은 각 인스턴스의 수행 종단에서 순간적으로 정보가 확보되는 상황에서 적용되어 왔으므로, 정보 발생의 환경이 바뀐 실시간 모

니터링에 적용될 때 그 효율이 반감됨을 알 수 있다. 이에 기인하여 다음과 같은 부수적인 한계점이 발생한다. 첫째, 인스턴스가 진행되는 동안 실시간으로 주도적인 대응을 할 수 없다. 기존의 접근법들은 프로세스의 종단에서 상태의 규명이 이루어지는 방식을 채택하고 있다[5]. 따라서, 각 인스턴스에 대한 상태의 규명은 프로세스가 완전히 종료되는 시점까지 연기될 수밖에 없으며, 그동안의 모니터링 시스템은 단순히 기다려야만 하는 유휴 상태에 놓이게 된다. 진행 중인 인스턴스가 이상치로 종료된다 하더라도, 이를 감지하는 시점은 완전히 종료되고 난 뒤이며 사용자에게 이상치의 발생을 알리고 그로 인한 손실을 보상(compensation) 또는 조정(correction)을 통해 해결해야할 뿐이다[3, 19]. 둘째, 실시간 진행 상태에 대한 직관을 사용자에게 전달하기 위한 정교하고 고도화된 지표가 없다. 룰 기반 모니터링은 인스턴스의 상태를 룰 조건부에 대한 만족 여부를 통해 확정적으로 정의한다[9]. 하지만, 진행 중인 인스턴스는 관측 시점 이후 다양한 경로로 진행될 수 있는 불확실성을 안고 있다. 따라서, 이러한 실시간 진행 경향성을 표현할 수 있는 지표가 필요하다[8]. 이러한 지표가 제공된다면, 내부 진행 상황에 대한 이해력을 높일 뿐 아니라, 종료 시점에서 원하는 관리 수준에 대한 만족 가능성을 실시간 진행에 걸쳐 주기적으로 가시화할 수 있다[18].

3. 연구 방법론

본 논문에서는 기존의 룰 기반 모니터링 방법론이 실시간 프로세스 모니터링에서 갖는 한계점을 해결하기 위하여 대체 기법을 도입하였다. 본 논문에서 새롭게 도입한 대체 기법에 대해 설명하고, 대체를 적용한 실시간 프로세스 모니터링 방법론의 주요 개념을 설명한다.

3.1 대체(Imputation) 기법

대체 기법이란 데이터의 결측치(missing value)가 발생한 경우, 이 값을 있음직한 특정 값으로 대체하기 위한 방법론이다[6]. 원자료(raw data)의 수집 과정에서 기술적인 문제나 인위적인 실수로 인해 데이터의 결측이 발생할 수 있다. 하지만 원자료를 특정 알고리즘이나 함수, 방법론의 입력 자료로 활용하기 위해서는 모든 데이터 테이블이 채워져야 할 필요가 있다. 이와 같이 불완전자료를 완전자료로 만들기 위해 활용되는 기법이 대체이다. 먼저, 데이터의 결측을 무시하는 방법이 있다. 전체 데이터 테이블에서 결측이 발생한 칸의 열 또는 행을 완전히 삭제함으로써 축소되었지만 완성된 데이터 테이블을 생성하는 기법이다. 이는 핵심 변수의 추출 과정을 거쳐 데이터의 차원을 축소시켜서 활용하는 경우 적용될 수 있지만, 정보의 손실로 인해 비교적 낮은 성능을 보인다. 따라서, 거리기반 유사도나 평균 등을 사용하여 특정 값을 채워 본래 데이터의 차원 감소나 정보 손실을 피하는 기법들이 널리 활용되고 있다.

본 논문에서 도입한 기법은 KNN(k nearest neighbor) 대체이다. 결측이 아닌 값들을 기준으로 유사도가 높은 샘플들을 선별하여 대체 값을 찾아내는 기법을 Hot Deck 대체

라 하며, KNN 대체는 가장 널리 활용되는 Hot Deck 대체 중 하나이다. 관측값들을 기준으로 거리 기반으로 유사한 이웃 샘플들을 추출하고, 이웃들이 가진 결측치의 평균으로 대체하는 것이다.

KNN 대체는 다음과 같이 이루어진다. 먼저, object x의 j번째($1 \le j \le n$) attribute가 결측되었음을 가정한다. 결측이 발생하지 않은 나머지 attributes를 이용하여 k개의 유사한 샘플들을 거리 기반으로 추출한다. 원자료의 샘플을 y, 속성들을 y_i라 할 때, 유사도는 다음과 같이 결측을 제외한 관측치를 이용한 거리로 계산된다.

$$d = \sqrt{\sum_{i=1}^{j-1}(x_i - y_i)^2 + \sum_{i=j+1}^{n}(x_i - y_i)^2} \qquad (2)$$

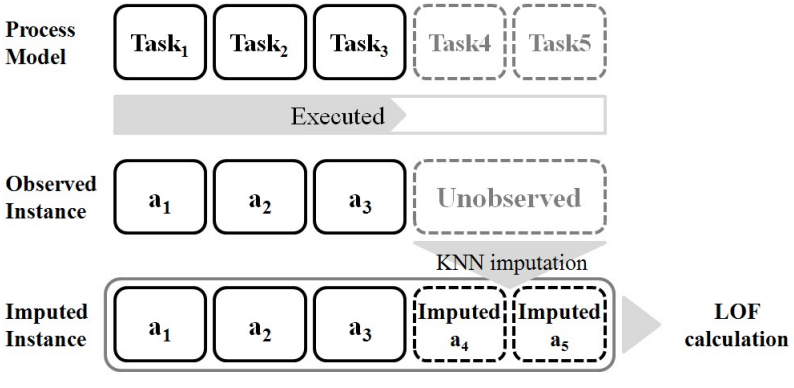
이 거리를 내림차순으로 정렬하고 상위 k개의 이웃들을 추출한 뒤, 이들이 가진 결측치의 평균으로 값을 대체한다. k번째 이웃 샘플 y_k가 가진 j번째 attribute를 $y_{j,k}$라 할 때, 결측에 대한 대체치는 다음과 같이 계산된다.

$$imputed\, x_j = \frac{\sum_{i=1}^{k} y_{j,i}}{k}$$

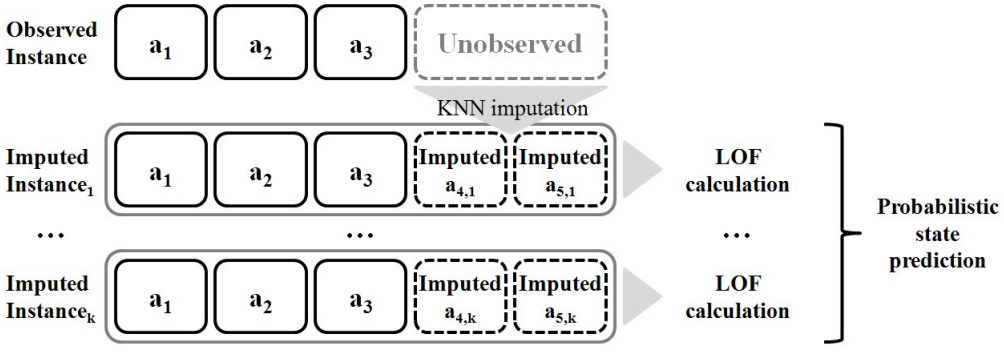
3.2 대체 기반 실시간 프로세스 모니터링

기존의 룰 기반 접근법들이 가진 한계점들의 근본적인 원인은 실시간 프로세스 진행 중 각 관측 시점에 따라 비관측 정보, 즉 관측 시점에 아직 수행되지 않은 프로세스 진행에 관련된 속성 정보가 획득되지 못함이다. 이러한 비관측정보의 존재를 대체를 도입함으로써 해결하고자 한다. 진행 중인 인스턴스의 속성 정보는 비완전 자료라 할 수 있으며, 대체를 통해 비획득 속성값이 가정함으로써 이를 완전자료화 할 수 있다. 아래 <그림 1>은 KNN 대체를 적용한 경우, 실시간 프로세스 모니터링이 수행되는 과정을 도식화하고 있다.

프로세스는 5개의 태스크(task)으로 구성되어 있고, 각 작업의 수행 시 관련 속성값이 하나씩 기록된다고 가정하였다. 3번째 태스크까지 수행된 시점에서의 진행 중인 인스턴



〈그림 1〉 KNN 대체 기반 실시간 프로세스 모니터링



〈그림 2〉 확장 KNNI 대체 기반 실시간 프로세스 모니터링

스가 관측된 상황에서 관측 인스턴스 로그는 완전자료가 아니므로 LOF를 적용하여 상태를 규명할 수 없다. 이때 비관측값을 KNN 대체를 통해 특정값으로 대체시킴으로써 진행 중인 인스턴스가 관측 시점 이후로 진행 가능한 상황을 가정하는 것이다. 관측된 속성값들과 대체된 속성값들의 집합은 하나의 완전자료를 구성함으로써 LOF 계산이 가능하기 때문에 일부 진행 시점에서의 인스턴스에 대한 상태의 규명이 가능하다. 하지만, 이와 같은 대체 기법을 단순히 적용한 대체 기반 모니터링 모델은 하나의 인스턴스 모델을 가정한다는 것이 한계점이 될 수 있다. 본 연구에서는 관측 시점 이후의 진행 가능한 다양한 결과들을 고려하는 것이 실시간 프로세스 모니터링의 궁극적인 목적에 부합한다고 도출하였다. 따라서, 하나의 대체 인스턴스를 생성하는 것이 아니라, 관측시점까지의 진행 경로에 근거한 다양한 대체 인스턴스들을 생성하기로 하였다. KNN의 평균값을 대체하는 것이 아니라, KNN 각각이 가진 값들을 대체하는 것이다. 따라서, k개의 대체 인스턴스를 다음과 같이 생성하기로 하였다. k번째 이웃 샘플 yk가 가진 j번째 attribute를 $y_{j,k}$라 할 때, 대체 인스턴스 i의 속성 정보는 다음과 같이 구성된다.

$$\{(\text{observed attributes}), (\text{imputed attributes})\} = \{(a_1, a_2, a_3), (y_{4,i}, y_{5,i})\}$$

KNN대체를 통해 k개의 대체 인스턴스를 생성함으로써, k개의 LOF를 계산할 수 있다. 이 값들을 이용하여 LOF 값의 분포를 가시화함으로써 다양한 결과들의 가능성에 대해 확률적으로 가시화하는 것을 본 논문이 제안하는 방법론의 최종목표로 한다. <그림 2>는 <그림 1>로부터 확장된 KNN 대체 기반 실시간 프로세스 모니터링 모델을 보여주고 있다.

4. 확장 KNNI 대체 기반 LOF 예측 알고리즘

본 절에서는 실시간 프로세스 모니터링을 위한 확장 KNNI 대체 기반 LOF 예측을 수

행하기 위한 알고리즘을 설명한다. 제안하는 알고리즘을 통하여 진행 중인 인스턴스의 일부 관측된 정보를 기반으로 다양한 대체 인스턴스들을 생성한다. 즉, 다양한 향후 진행 경로들을 가정하고 각 경우들이 갖는 LOF를 계산한 후, 이 값들이 갖는 확률 분포를 가시화함으로써 향후 진행 경로에 대한 다양성 및 불확실성을 함께 추후 상태를 예측할 수 있도록 하는 것이 본 논문에서 제안하는 알고리즘의 목적이다. 특정 관측 시점에 관측된 실시간으로 진행 중인 인스턴스의 속성 정보를 입력으로 하여 알고리즘은 다음 단계들을 거쳐 수행되고 해당 인스턴스가 종료된 시점에서 갖게 될 LOF의 확률분포함수를 출력한다.

(Step 1) 진행 중인 인스턴스의 관측

프로세스가 실행되면 m개의 attributes로 구성된 인스턴스 로그가 관측된다. 모니터링 시점마다 하나의 attribute가 추가로 관측됨을 가정한다. 따라서, 모니터링 시점 t에 시스템에는 (a_1, …, a_t)의 인스턴스 로그가 관측된다. 이를 Observation at t, 즉 O_t라 정의한다.

(Step 2) 부분 정보 기반 KNN 추출

관측된 일부 속성들을 기준으로 식 (2)의 거리 기반 유사도를 과거 사례 각각에 대하여 계산한다. 그리고 유사도가 높은 상위 k개의 과거 사례들을 추출한다.

(Step 3) K 대체 인스턴스 생성

확장 KNNI 대체를 이용하여 k개의 대체 인스턴스를 생성한다. O_t의 i번째 대체 인스턴스 Imputed_$O_{t,i}$는 다음과 같이 구성된다.

{(observed attributes), (imputed attributes from i-th nearest historical instance)} = {(a_1,…, a_t), ($a_{t+1,i}$,…, $a_{m,i}$))

이때, $a_{j,i}$는 i번째 이웃 인스턴스가 가진 j번째 attribute이다.

(Step 4) 각 대체 인스턴스의 LOF 계산

각 Imputed_$O_{t,i}$에 대해 식 (1)의 LOF 알고리즘을 수행하여 LOF를 계산한다. 이 값을 LOF(Imputed_$O_{t,i}$)라 정의한다.

(Step 5) LOF들의 확률분포 계산

k개의 LOF들이 각각 별개로 계산되었다. 이들을 핵밀도추정(kernal density estimation, KDE)을 이용하여 확률분포함수(probabilistic distribution function)를 생성한다. 목적 변수를 LOF(O_t), 즉 Ot가 종료된 시점에서 갖게 될 LOF라 할 때, 그 값의 확률분포함수는 아래 식을 따른다.

$$f(x) = \frac{1}{k}\sum_{i=1}^{k} K(\frac{x - x_i}{h})$$

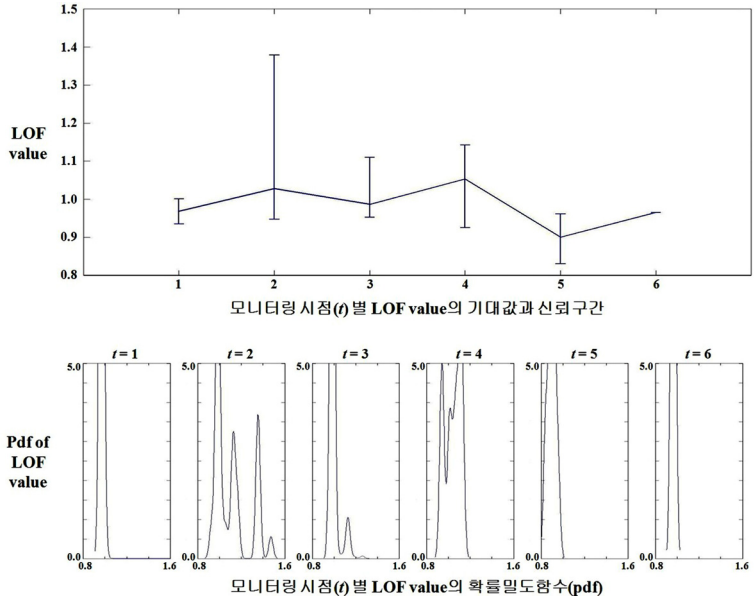
이때, x는 목적변수인 LOF(O_t)가 되고, xi는 각 LOF(Imputed_$O_{t,i}$)에 해당된다. K는 밀도추정에 활용되는 kernel 함수를 의미하며, 표준정규분포를 따르며 아래 식과 같다.

$$K(\frac{x - x_i}{h}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$

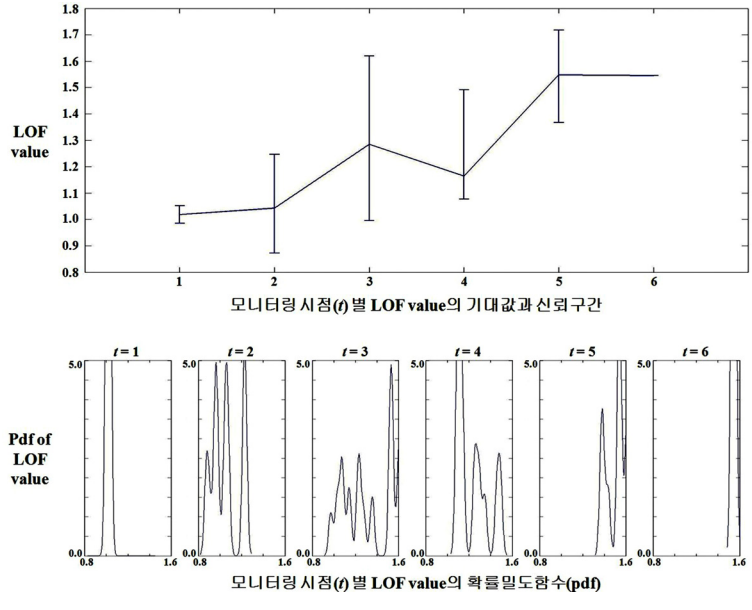
h는 평활계수(smoothing parameter)로 불리며, 각 kernel의 분산으로 정의된다. 본 논문의 실험에서는 matlab을 이용하여 자동으

로 최적화된 h값을 찾아주는 프로그램을 사용하였다.

위에서 제시된 전 과정을 걸쳐, 각 모니터링 시점에서 관측된 일부 정보에 근거하여



〈그림 3〉 정상 종료 인스턴스의 실시간 모니터링 결과



〈그림 4〉 비정상 종료 인스턴스의 실시간 모니터링 결과

해당 인스턴스가 종료시점까지 진행되었을 때 예측되는 LOF의 분포를 가시화할 수 있다. 분포함수를 가시화하는 이유는 다음과 같다. 첫째, 평균값을 통해 전반적인 추세를 알 수 있다. 둘째, 분산을 통해 불확실성에 근거한 분포 상태를 알 수 있다. 셋째, 양태(modality)를 통해 분포의 치우침을 알 수 있다. 같은 평균값을 갖더라도 multi modality를 가질 경우, 진행 중인 인스턴스는 각 modality의 방향성을 독립적으로 가짐을 알 수 있기 때문이다.

5. 실시간 프로세스 모니터링 예제

본 논문에서 제안하는 확장 KNNI 대체 기반 LOF 알고리즘 및 이를 활용한 비정상적 종료 예측 기반 실시간 프로세스 모니터링 방법론의 효용성을 입증하기 위하여, 다음과 같은 예제 시나리오를 이용한 실험을 수행하였다.

본 실험의 목적은 다음과 같다. 제안된 알고리즘을 실시간 모니터링에 적용하여 진행중인 프로세스 인스턴스가 얼마나 비정상적으로 진행되고 있는지를 LOF 값의 기댓값, 신뢰구간, pdf를 통해 정량적으로 가시할 수 있다. 완전히 종료된 인스턴스의 속성값으로 구성된 벡터로부터 계산된 LOF value는 해당 인스턴스가 얼마나 비정상적으로 종료되었는지를 의미한다. LOF value가 1을 크게 초과할 경우, 해당 인스턴스는 과거에 수행된 로그들에 비하여 비정상적으로 수행된 경우이므로, 그에 상응하는 조치가 요구된다. 하지만 이러한 LOF value의 계산은 인스턴스가 완전히 종료된 후에만 가능하다. 하지만 제안된 알고리즘을 통하여 인스턴스가 종료되기 전에 일부 관측 정보만으로 종료된 시점에서의 LOF의 pdf를 예측할 수 있다. 이 예측된 pdf는 진행 중인 인스턴스가 완전히 종료되었을 때 가질 수 있는 LOF 값들의 분포를 의미한다. 따라서, 전 관측 시점에 걸쳐 종료 시점에서의 LOF의 pdf, 기댓값, 신뢰구간을 살펴봄으로써 비정상적인 종료를 사전에 예측할 수 있다. 비정상적으로 종료되는 인스턴스의 경우, 시점에 걸쳐 LOF의 기댓값 및 신뢰구간이 점차적으로 1보다 높은 값으로 상승하게 될 것이라 예상할 수 있다. 반대로 정상 종료의 인스턴스는 1보다 낮은 값 또는 1에 수렴하는 LOF 값을 갖게 되고, 그에 따라 관측 시점에 걸쳐 하락할 것으로 예측된다.

프로세스 모델은 6개의 태스크가 순차적으로 수행되며, 각 태스크가 수행되면 하나의 속성이 관측 및 기록된다. 따라서, 6회의 관측시점이 존재하며 마지막 6번째 관측시점은 프로세스의 종료를 의미한다. 즉, 6번째 관측 시점에서의 LOF는 참값이며 실제 진행 중인 인스턴스가 완료된 시점에서의 LOF를 의미한다. 10000개의 인스턴스는 3개의 속성은 정규분포로, 3개의 속성은 가우시안 혼합(Gaussian Mixture) 분포로 무작위 생성하였다. 또한 6개의 속성들은 각각의 독립적이고 동등한 영향력을 가지도록 하기 위해 랜덤 생성 후 정규화 하였다. 10000개의 인스턴스들 중, 8000개는 과거 인스턴스 로그로 사용되었고, 2000개의 인스턴스는 테스트, 즉 실시간 모니터링 샘플로 활용하였다.

진행 중인 인스턴스를 실시간 모니터링하

는 과정에서 가시화되는 정보를 두 가지 예제 샘플을 통하여 살펴본다. 제안된 알고리즘을 통해 진행 중인 인스턴스가 종료 시점에서 갖게 될 LOF를 확률분포함수로 예측한다. <그림 3>에서는 정상으로 종료되는 인스턴스를 모니터링 하는 과정에서 획득된 정량화된 지표들을 보여준다. <그림 3>의 (상)은 각 시점 별 분포함수로부터 LOF의 평균값과 90% 신뢰구간을 추출하여 모니터링 시점에 걸친 예측값들의 추세를 보여준다. 그리고 <그림 3>의 (하)는 각 시점별 LOF의 확률분포함수이다. 동일하게, 비정상적으로 종료되는 인스턴스에 대한 실시간 모니터링 결과를 <그림 4>에서 보여주고 있다.

실험 결과, 다음과 같은 특징들을 분석할 수 있었다. 첫째, 모니터링 시점이 진행될수록 분포함수의 modality 수가 감소하였다. 다수의 modality를 가질 경우 독립된 다수개의 패턴으로 진행될 가능성을 각각 가지고 있음을 의미하지만, 단일화가 진행될수록 예측된 결과들이 단일 패턴으로 통합된다. 둘째, 종료결과에 따라 LOF의 평균값은 일정한 추세를 가짐을 알 수 있었다. 비정상적으로 종료되는 인스턴스의 경우, 시점이 경과될수록 신뢰구간의 감소와 함께 평균값의 상승이 나타났다. LOF는 일반적으로 이상치의 경우 1보다 높은 값을 갖게 되며, 정상의 경우 1에 수렴하게 된다. 이와 같은 특성이 반영되어, 비정상 종료 인스턴스는 종료 시점에서 그에 상응하는 1을 초과한 LOF를 가졌고, 종료 시점이 가까워질수록 그에 수렴하며 점진적으로 상승하였다. 그와 반대로 정상 종료 인스턴스는 점차 낮아지는 추세를 보였다. 두 경우 모두 초반부의 시점에서는 그 중간에 해당하는 불확실한 LOF 값과 넓은 신뢰구간을 가졌다.

본 실험과 종료 타입 별 예제를 통해 본 논문이 제안하는 알고리즘을 이용한 실시간 모니터링 방법론으로 인스턴스의 실시간 진행 상황이 어떻게 지표화되어 가시화되는지 살펴보았다. 전 관측 시점에 걸쳐 종료 타입에 따른 추세가 나타남을 알 수 있었으며 또한 시점이 경과될수록 그 값의 정확도가 높아졌다. 따라서, LOF 값의 확률분포함수와 평균값, 신뢰구간이라는 지표들을 통해 진행 중인 인스턴스의 종료 패턴에 대한 예측이 가능하였다.

6. 결 론

본 논문에서는 비정상적인 종료의 예측을 위한 실시간 프로세스 모니터링 방법론을 제안하였다. 또한 이를 위해 KNN 대체 기반 LOF 예측 알고리즘을 제안하였다. 기존의 룰 기반 접근법들이 실시간 프로세스 모니터링에 적용될 경우, 관측 시점에 따른 비획득 정보의 발생에 기인하여 실시간 진행 상태를 가시화하기 위한 지표가 부재하며 비정상 종료의 발생을 조기에 감지할 수 없다는 한계점들이 발생하였다. 이를 해결하기 위하여 KNN 대체 기법을 새롭게 도입하였으며, 실시간 모니터링의 목적에 부합할 수 있도록 보다 다양한 결과들을 확률적으로 예측하기 위하여 이를 확장하였다. 실시간으로 진행 중인 인스턴스가 가진 부분 정보를 기반으로 대체를 통해 다수 개의 완성된 대체 인스턴스를 가정함으로써, 인스턴스가 계속적인 진

행을 통해 종료 시점에서 갖게 될 패턴들을 예측하였다. 완성된 인스턴스 로그를 가정함으로써 기존의 룰 기반 접근법들이 실시간 환경에서도 적용될 수 있었다. 제안된 KNN 대체 기반 LOF 예측 알고리즘을 이용하여 진행 중인 인스턴스가 종료 시점에서 갖게 될 LOF의 확률분포함수를 추정할 수 있었다. 이는 매 관측시점마다 수행되며 전 관측 시점에 걸친 지표들의 변화과정을 통해 인스턴스의 종료 상황을 예측할 수 있다.

제안된 기법을 프로세스 모니터링에 적용할 경우, 기존의 접근법들이 제시할 수 없었던 실시간 진행 상태를 나타내는 지표를 산출할 수 있다. 이를 통해 보다 능동적인 대응을 위한 지표로서 활용이 가능할 것이다. 기존의 인스턴스 단위의 관리 수준에서 개별 인스턴스의 전 진행 시점으로 모니터링의 관점을 전환시킴으로써, 프로세스 관리자에게 프로세스 내부의 실시간 진척 상황에 대한 정보를 더욱 구체적으로 제공할 수 있을 것이다. 더욱이, 시점에 걸쳐 종료 패턴에 대한 직관을 줄 수 있는 패턴을 보여줄 수 있기 때문에, 이는 위협 및 기회에 대해 보다 능동적으로 대응할 수 있는 근거를 마련함으로써 관리 수준의 비약적인 향상을 기대할 수 있을 것이다.

추후 연구 과제로는 결과 지표들에 기반한 조기 경보 전략의 수행이다. 대응적 경보의 수준은 이미 일어난 비정상 종료에 대한 보상 및 조정을 위한 단서를 제공할 뿐이다. 제안된 알고리즘을 통해 종료 패턴을 사전에 가시화할 수 있으므로 비정상 종료를 조기에 감지함으로써 사전에 경보를 내리고 비정상 종료의 실질적인 발생을 방지하기 위한 대응책을 수행할 수 있을 것이다. LOF 확률분포에 대한 상위한계값의 결정하고 이를 초과할 경우 조기에 경보를 내림으로써, 추후의 진행 과정동안 추가적인 리소스의 투입이나, 보상을 위한 메타 프로세스의 호출 등을 통해 비정상 종료를 방지할 수 있다. 이를 위해서는 조기 경보의 정확도 및 그에 따른 손익 분석을 통해 최적의 조기 경보 수준을 위한 LOF 한계값 결정이 필수적이다.

참 고 문 헌

[1] Breunig, M. M., Kriegel, H. P., Ng R. T. and Sander, J., “LOF : Identifying Density Based Local Outliers,” In Proceedings of the ACM SIGMOD Conference, Dallas, TX, 2000.

[2] Buytendijk, F. and Flint, D., “How BAM can turn a business into a real-time enterprise,” Gartner Research, AV-15-4650, 2002.

[3] Castellanos, M., Salazar, N., Casati, F., Dayal U. and Shan, M. C., “Predictive business operations management,” International Journal of Computational Science and Engineering, Vol. 2, No. 5/6, 2006, pp. 292-301.

[4] Chen, J. C. and Lin, K. Y., “Diagnosis for monitoring system of municipal solid waste incineration plant,” Expert Systems with Applications, Vol. 34, No. 1, 2008, pp. 247-255.

[5] Curtis, B., Seshagiri, G. V., Reifer, D., Hirmanpour, I. and Keeni, G., “The cases

for quantitative process management," IEEE software, Vol. 25, No. 3, 2008, pp. 24-28.

[6] Ek, A. R., Robinson, A. P., Radtke, P. J. and Walters, D. K., "Development and testing of regeneration imputation models for forests in Minnesota," Forest Ecology and Management, Vol. 94, No. 1-3, 1997, pp. 129-140.

[7] Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal M., and Shan, M. C., "Business process intelligence," Computers in Industry, Vol. 53, No. 3, 2004, pp. 321-343.

[8] Grigori, D., Casati, F., Dayal U., and Shan, M. C., "Improving business process quality through exception understanding, prediction, and prevention," In Proceedings of the 27th Very Large Data Base Endowment Conference, Roma, Italy, 2001, pp. 159-168.

[9] Kang, B., Cho N. W. and Kang, S. H., "Real-time risk measurement for Business Activity Monitoring(BAM)," International Journal of Innovative Computing, Information and Control, Vol. 5, No. 11(A), 2009, pp. 3647-3657.

[10] Kang, B., Lee, S. K., Min, Y., Kang S. H. and Cho, N. W., "Real-time Process Quality Control for Business Activity Monitoring," In Proceedings of the 2009 International Conference on Computational Science and Its Applications, Yongin, Korea, 2009, pp. 237-242.

[11] Keung, P. and Kawalek, P., "Goal-based business process models: Creation and evaluation," Business Process Management Journal, Vol. 3 No. 1, 1997, pp. 17-38.

[12] Kim, K., Choi, I. and Park, C., "A rule-based approach to proactive exception handling in business processes," Expert Systems with Applications, Vol. 38, No. 1, 2011, pp. 394-409.

[13] Lazarevic, A., Ertoz, L., Ozgur, A., Srivastava, J. and Kumar, V., "A comparative study of anomaly detection schemes in network intrusion detection," In Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, 2003.

[14] Leitner, P., Wetzstein, B., Rosenberg, F., Michlmayr, A., Dustdar S., and Leymann, F., "Runtime prediction of service level agreement violations for composite services," In Proceedings of the 3rd Workshop on Non-Functional Properties and SLA Management in Service-Oriented Computing, Stockholm, Sweden, 2009.

[15] Medioni, G., Cohen, I., Hongeng, S., Bremond F., and Nevatia. R., "Event Detection and Analysis from Video Streams," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 8, No. 23, 2001, pp. 873-889.

[16] Nie, G., Zhang, L., Liu, Y., Zheng, X. and Shi, Y., "Decision analysis of data mining project based on Bayesian risk," Expert Systems with Applications, Vol. 36, No. 3, 2009, pp. 4589-4594.

[17] Pokrajac, D., Lazarevic, A. and Latecki, L. J., "Incremental Local Outlier Detection for Data Streams," IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2007.

[18] Rao, U. S., Kestur, S. and Pradhan, C., "Stochastic optimization modeling and quantitative project management," IEEE software, Vol. 25, No. 3, 2008, pp. 29-36.

[19] Rusinov, L. A., Rudakova, I. V., and Kurkina, V. V., "Real time diagnostics of technological processes and field equipment," Chemometrics and Intelligent Laboratory Systems, Vol. 88, No. 1, 2007, pp. 18-25.

[20] Viaene, S., Dedene, G. and Derrig, R. A., "Auto claim fraud detection using Bayesian learning neural networks," Expert Systems with Applications, Vol. 29, No. 3, 2005, pp. 653-666.

[21] Wang, D. and Romagnoli, J. A., "Robust multi-scale principal components analysis with applications to process monitoring," Journal of Process Control, Vol. 15, No. 8, 2005, pp. 869-882.

[22] Yue, D., Wu, X., Wang, Y., Li, Y. and Chu, C. H., "A Review of Data Mining-Based Financial Fraud Detection Research," In Proceedings of 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China, 2007, pp. 5514-5517.

저 자 소 개

강복영	(E-mail : realgas@snu.ac.kr)
2005년	한국과학기술원 산업공학과 (학사)
2007년	서울대학교 산업공학과 (석사)
2007년~현재	서울대학교 산업공학과 (박사과정)
관심분야	기업정보시스템, BPM, BAM(Business Activity Monitoring), Real-time System, Data Mining

김동수	(E-mail : dskim@ssu.ac.kr)
1994년	서울대학교 산업공학과 (학사)
1996년	서울대학교 산업공학과 (석사)
2001년	서울대학교 산업공학과 (박사)
2001년~2003년	한국정보사회진흥원 전자거래연구부 e-Biz 표준팀장
2003년~2006년	가톨릭대학교 의료경영대학원 전임강사, 조교수
2006년~현재	숭실대학교 산업·정보시스템공학과 조교수, 부교수
관심분야	BPM, e-Business 정책 및 기술, 기업정보시스템, e-Health

강석호	(E-mail : shkang@snu.ac.kr)
1970년	서울대학교 물리학과 (학사)
1972년	University of Washington 산업공학과 (석사)
1976년	Texas A&M University 산업공학과 (박사)
1976년~1987년	서울대학교 산업공학과 조교수, 부교수
1987년~현재	서울대학교 산업공학과 교수
관심분야	Intelligent Manufacturing System, Management Information System, Business Process Management