KIPS Trans. Softw. and Data Eng.
Vol.7, No.11 pp.435~442 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2018.7.11.435

Outlier Detection By Clustering-Based Ensemble Model Construction

Cheong Hee Park† · Taegong Kim†† · Jiil Kim†† · Semok Choi†† · Gyeong-Hoon Lee††

ABSTRACT

Outlier detection means to detect data samples that deviate significantly from the distribution of normal data. Most outlier detection methods calculate an outlier score that indicates the extent to which a data sample is out of normal state and determine it to be an outlier when its outlier score is above a given threshold. However, since the range of an outlier score is different for each data and the outliers exist at a smaller ratio than the normal data, it is very difficult to determine the threshold value for an outlier score. Further, in an actual situation, it is not easy to acquire data including a sufficient amount of outliers available for learning. In this paper, we propose a clustering-based outlier detection method by constructing a model representing a normal data region using only normal data and performing binary classification of outliers and normal data for new data samples. Then, by dividing the given normal data into chunks, and constructing a clustering model for each chunk, we expand it to the ensemble method combining the decision by the models and apply it to the streaming data with dynamic changes. Experimental results using real data and artificial data show high performance of the proposed method.

Keywords : Streaming Data, Ensemble Method, Outlier Detection, K-Means Clustering

클러스터링 기반 앙상블 모델 구성을 이용한 이상치 탐지

박 정 희† · 김 태 공†† · 김 지 일†† · 최 세 목†† · 이 경 훈††

요 약

이상치 탐지는 정상 데이터 분포를 크게 벗어나는 데이터 샘플을 탐지하는 것을 의미한다. 대부분의 이상치 탐지 방법은 데이터 샘플이 정상 상태를 벗어나는 정도를 나타내는 이상치 지수(outlier score)를 계산하여 주어진 임계값 이상일 때 이상치로 판정한다. 그러나, 데이터마다 이상치 지수의 범위가 다양하고 정상 데이터에 비해 이상치 데이터는 적은 비율로 존재하기 때문에 이상치 지수에 대한 임계값을 결정하기는 매우 어렵다. 또한, 실제 상황에서는 학습에 이용할 수 있는 충분한 양의 이상치를 포함하는 데이터의 획득이 용이하지 않다. 본 논문에서는 정상 데이터가 주어졌을 때 이를 이용하여 정상 데이터 영역을 나타내는 모델을 구성하고 새로운 데이터 샘플에 대해 이상치와 정상치의 이진 분류를 수행하는 방법으로 군집화 기반 이상치 탐지 방법을 제안한다. 그리고, 주어진 정상 데이터를 청크로 나누고 각 청크에 대해 클러스터링 모델을 구성한 후 모델들에 의한 이상치 판정 결과를 결합하는 앙상블 방법과 동적 변화가 있는 스트리밍 데이터에서의 적용 방법으로 확장한다. 실제 데이터와 인공 데이터를 이용한 실험결과는 제안 방법의 높은 성능을 보여준다.

키워드 : 스트리밍 데이터, 앙상블 방법, 이상치 탐지, K-Means Clustering

1. 서 론

이상치 탐지는 정상 데이터 분포를 크게 벗어나는 데이터 샘플을 탐지하는 것을 의미한다[1]. 이상치 탐지 방법을 이용하여 설비운용에서 이상 상태 감지나 네트워크 침입 탐지 등을 실시간으로 수행할 수 있게 된다. 거리 기반, 밀도 기반, 클러스터링 기반, 트리 기반 방법 등 다양한 이상치 탐지 방법이 개발되어 왔다[2-5].

대부분의 이상치 탐지 방법은 정상 상태를 벗어나는 정도를 나타내는 이상치 지수(outlier score)를 계산하는 방법을 제시한다. 이상치 지수가 상위 k번째 이내에 위치한 데이터 샘플들을 이상치로 판정하거나 주어진 임계값보다 큰 이상치 지수를 가지는 데이터 샘플을 이상치로 판정한다. 그러나, 대부분의 경우 정상 데이터에 비해 이상치 데이터는 거의 없거나 매우 적은 비율로 존재하기 때문에 이상치 지수에 대한

※ 본 연구는 한국전력공사의 2018년 착수 에너지 거점대학 클러스터 사업에 의해 지원되었음(과제번호: R18XA05).
† 정 회 원 : 충남대학교 컴퓨터공학과 교수
†† 비 회 원 : 충남대학교 컴퓨터공학과 석사과정
Manuscript Received : July 26, 2018
First Revision : August 27, 2018
Accepted : September 17, 2018
* Corresponding Author : Cheong Hee Park(cheonghee@cnu.ac.kr)

임계값을 결정하기는 매우 어렵다[5].

정상 데이터와 이상치 데이터가 함께 존재하는 학습데이터가 주어질 때 이진분류기를 학습하여 이상치와 정상치를 분류하는 감독학습 방법을 적용할 수 있으나, 실제 상황에서는 학습에 이용할 수 있는 충분한 양의 이상치를 포함하는 데이터의 획득이 용이하지 않다는 문제가 있다. 이러한 경우 정상 데이터만으로 정상 상태를 나타내는 모델을 구성하고 모델에서 벗어나는 정도를 이용하여 이상치를 탐지하는 학습 방법을 사용할 수 있다[5, 6].

대부분의 이상치 탐지 방법은 이상치가 섞여 있는 전체 데이터에 대해 이상치 지수를 계산하고 임계값 이상의 이상치 지수를 가지는 데이터 샘플들을 이상치로 예측한다. 본 논문에서는 정상 데이터로 모델을 먼저 구성하고 이를 이용해 이상치를 탐지하는 방법을 제안한다. 정상 데이터 집합에 대해 k-means clustering을 적용하여 정상 데이터 영역을 설정하고 이 영역을 벗어나는 데이터를 이상치 데이터로 판정한다. 그러나 초기 중심점들과 k값의 선택에 따라 k-means clustering의 결과는 다양하게 나타날 수 있어 학습되는 정상 데이터 영역 모델에 대한 신뢰성이 낮아질 수 있다. 따라서 주어진 정상 데이터를 청크로 나누고 각 청크에 대해 클러스터링 모델을 구성한 후 모델들에 의한 이상치 판정 결과를 결합하는 앙상블 방법을 수행한다. 또한, 데이터 생성 분포의 동적 변화 가능성이 있는 스트리밍 데이터에서 이상치 탐지를 위한 방법으로 확장한다.

제안 방법은 정상 영역 모델 구성을 위해 정상 데이터가 학습데이터로 주어진다는 전제를 바탕으로 하고 있다. 이상현상이 발생하지 않은 정상 상태에서의 데이터의 수집은 상대적으로 용이하다는 점과 이상치 여부에 대한 true/false 예측값을 주는 제안 방법의 특징은 설비운용에서의 이상 현상 발생 탐지와 같은 다양한 실제 문제에의 적용을 용이하게 할 수 있다.

2절에서는 이상치 탐지에 관련된 기존연구를 살펴보고 3절에서 클러스터링 기반 앙상블 모델에 의한 이상치 탐지 방법을 제안한다. 4절에서 동적 변화가 있는 스트리밍 데이터에서 앙상블 모델 업데이트에 의한 이상치 탐지를 설명한다. 5절에서 실험결과를 나타내고 6절에서 결론을 맺는다.

2. 관련 연구

이상치 탐지는 다른 대부분의 데이터와는 크게 다른 데이터 샘플을 탐지하는 것을 의미한다. 대부분의 데이터는 정상 데이터인 반면 정상을 크게 벗어나는 이상치는 소수일 것이라는 전제를 바탕으로 한다. 이상치 탐지 알고리즘의 출력값은 다음 두 가지 타입 중의 하나일 수 있다[5].

* 이상치 지수(Outlier score): 각 데이터 샘플에 대해 이상치일 가능성을 수치화해서 나타낸 값이다. 이상치 지수에 따라 데이터 샘플들을 정렬할 수 있고 상위에 랭크된 데이터 샘플들을 이상치로 예측할 수 있다.

* 이진 레이블(Binary label): 데이터 샘플이 정상치인지 이상치인지를 나타내는 이진 레이블을 출력으로 한다. 실제 응용 문제에서 의사 결정에 필요한 결과값일 수 있다.

다양한 이상치 탐지 방법들이 제안되어 왔으며, 대략적으로 거리 기반 방법, 밀도 기반 방법, 트리 기반 방법, 클러스터링 기반 방법, 딥러닝 기반 방법 등의 카테고리로 나눌 수 있다. 각 카테고리에 속하는 대표적인 이상치 탐지 방법을 살펴본다.

논문 [1]에서는 데이터 샘플의 R 반경 안에 k보다 적은 수의 데이터들이 있을 때 이상치로 판정하는 거리 기반 이상치 탐지 방법이 제안되었다. 이를 확장하여 이상치에 대한 이진 판정 대신 이웃들과의 거리를 이용하여 이상치 지수를 계산하기도 한다[7]. 그러나 거리 기반 방법은 두 변수 k와 R의 선택에 매우 민감하고, 서로 다른 밀도의 영역이 존재할 때 두 영역에 모두 적합한 변수 값의 설정이 어렵다는 단점이 있다.

이러한 단점을 극복하는 방법으로 데이터 샘플 주위의 상대 밀도에 기반한 이상치 지수인 LOF(Local outlier factor)가 제안되었다[8]. 한 데이터 샘플 주변 밀도가 k개의 근접한 이웃들의 주변 밀도 평균에 비해 낮을수록 높은 LOF를 가지게 된다. 그러나 새로운 데이터 샘플이 들어올 때마다 기존 데이터와 새로운 데이터에 대해 k개의 근접 이웃 계산이 업데이트 되어야 해서 시간과 공간 복잡도가 높다.

트리 기반 방법인 Isolation Forest[9]는 이상치는 정상 데이터보다 주위에서 고립되기 쉽다는 전제 하에 모든 데이터 샘플들이 분리될 때까지 반복해서 특징값과 분할점을 랜덤하게 선택하여 Isolation Tree를 구성한다. 학습데이터의 랜덤 샘플링에 의해 Isolation Tree들의 앙상블을 구성하고, 데이터 샘플이 루트로부터 리프노드에 도달할 때까지 경로의 길이를 이용해서 이상치 지수를 계산한다.

최근에는 딥러닝 모델을 이용한 이상치 탐지 방법이 제안되었다. 입력 데이터에 대해 출력값이 입력과 최대한 가까워지도록 신경망을 훈련하여 데이터 샘플의 출력 오차를 이용한 이상치 지수를 계산한다[10]. 이상치 지수는 정상 데이터 범주에서 벗어나 있는 정도를 수치로 제공해 주는 장점이 있지만, 반면에 이상치 지수값의 범위가 데이터마다 달라 이상치 판단을 위한 이상치 지수의 임계값 설정이 어렵다는 단점이 있다.

클러스터링 기반 이상치 탐지 방법은 학습데이터에 클러스터링 기법을 적용하여 작은 사이즈의 클러스터에 속하는 데이터를 이상치로 간주한다. [11]에서는 데이터 샘플이 속하는 클러스터의 크기와 클러스터 센터까지의 거리를 이용하여 지역적 이상치 지수(local outlier factor)를 계산한다. [12]에서는 정상 데이터로 구성된 학습데이터에 대해 k 개의 클러스터를 생성한 후 스트리밍 모드로 새로운 데이터가 입력될 때 가장 가까운 클러스터의 영역을 벗어나는 경우 “unknown profile”이라고 선언되고 이러한 샘플들에 대해 클러스터링을 다시 하여 기존 모델의 확장인지 새로운 개념의 생성인지 결정하여 모델을 갱신해 나간다.

우리는 [12]에서의 방법을 확장하여 정상 데이터를 청크로 분할하고 각 청크에 대해 k-means clustering을 적용하여 정상 모델들의 앙상블을 구성한다. 모든 데이터가 한꺼번에 주어지는 배치 모드에서 뿐만 아니라 데이터 분포가 시간에 따라 변할 수 있는 스트리밍 모드에서의 적용 방법도 제안한다.

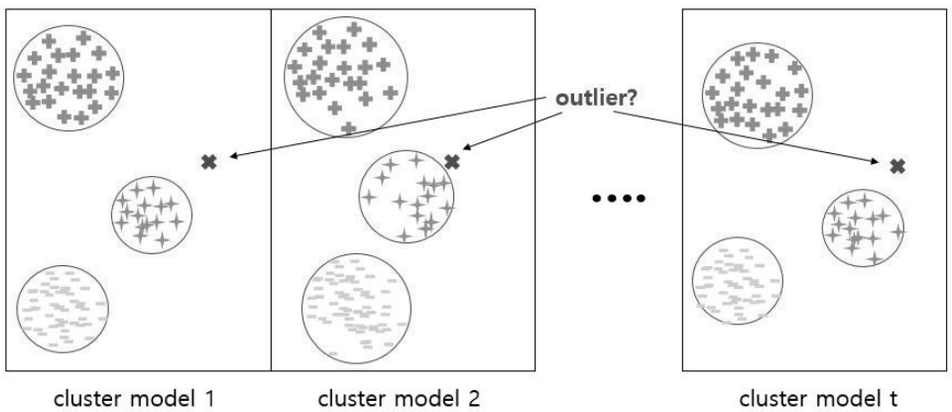


Fig. 1. An Illustration of a Clustering-Based Ensemble Method for Outlier Detection

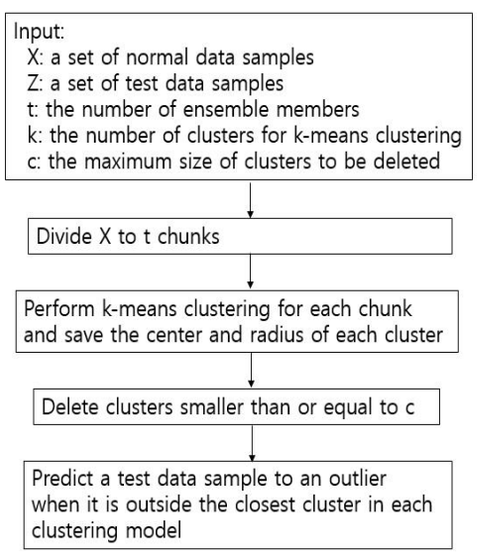


Fig. 2. The Process of a Clustering-Based Ensemble Method for Outlier Detection

3. 클러스터링 기반 이상치 탐지 방법

Fig. 1은 클러스터링 기반 이상치 탐지 방법을 보여준다. 정상 데이터 셑을 t개의 청크로 나눈 후 각 청크에 대해 군집화를 수행하여 정상 영역을 모델링한다. 정상 데이터의 영역을 클러스터 영역들의 합집합으로 나타낼 수 있다. 정상 데이터의 클러스터링에 의한 모델 구성 후에 새로운 데이터 샘플에 대해 각 청크의 클러스터링 모델에서 벗어나는지 여부를 체크한다. 모든 클러스터링 모델에서 이상치로 예측될 때 데이터 샘플은 최종적으로 이상치로 판정된다. Fig. 2에서 클러스터링 기반 앙상블 이상치 탐지 방법의 과정들을 요약하고 다음에서 각 과정을 상세히 설명한다.

3.1 클러스터링에 의한 정상 영역 모델링

주어진 정상 데이터를 t개의 청크로 나누고, 각 청크에 대해 클러스터링을 수행한다. 클러스터링 방법은 어느 것을 사용해도 되나 우리는 k-means clustering을 이용하였다. 다른 클러스터링 방법에 비해 k-means clustering에 의해 얻어지는 클러스터 영역은 센터와 반경에 의해 나타냄으로써 데이터 샘플이 클러스터 영역 내에 있는지 판단하기 쉽기 때문이다. 클러스터 반경은 센터로부터 클러스터 멤버들까지의 거리 중에서 가장 먼 거리를 의미한다. k-means clustering에서 클러스터 개수를 나타내는 k값은 정상 데이터를 둘러싸는 정상영역이 어떻게 구성되는가에 영향을 미치므로 아웃라이어 판정 정확성도 달라지게 된다. 너무 작은 k값을 이용하면 적은 수의 hypersphere 안에 모든 정상 데이터를 담아야 하므로 클러스터 반경이 커지게 되고 아웃라이어를 정상 데이터로 판단하게 되는 FN(false negative) 비율이 높아진다. 반면에 k값이 너무 커지면 정상영역을 타이트하게 설정하게 되어 정상 데이터를 아웃라이어로 판정하는 FP(false positive) 비율이 높아진다.

3.2 작은 사이즈의 클러스터 삭제에 의한 노이즈 제거

정상 데이터에 존재하는 노이즈는 클러스터링 결과에도 영향을 미치지만 클러스터 반경을 크게 함으로써 정상영역 설정에도 영향을 미친다. k-means 클러스터링에서 다소 큰 k값을 사용함으로써 대부분의 정상 데이터로부터 노이즈를 포함하는 작은 사이즈의 클러스터를 분리하여 구성할 수 있다. 노이즈로 구성된 작은 사이즈의 클러스터를 정상 모델로부터 제거함으로써 정상영역을 효율적으로 구성할 수 있게 된다.

3.3 앙상블 모델에서의 이상치 판정

새로운 데이터 샘플에 대해 이상치 여부를 판정하기 위해 각 청크의 클러스터링 모델마다 가장 가까운 클러스터 센터를 찾아 클러스터 반경을 벗어나는지 확인한다. 모든 앙상블 멤버에서 정상영역을 벗어날 때 데이터 샘플은 이상치로 판정된다. 모든 앙상블 멤버가 이상치로 일치된 의견을 낼 때에 이상치로 판정하는 것은 정상 데이터 영역을 각 앙상블 멤버에 의한 정상영역의 합집합으로 간주하는 것이다. 따라서 앙상블 멤버 수가 너무 많아지면 이상치를 정상으로 판정하는 FN(False negative)가 커질 수 있다.

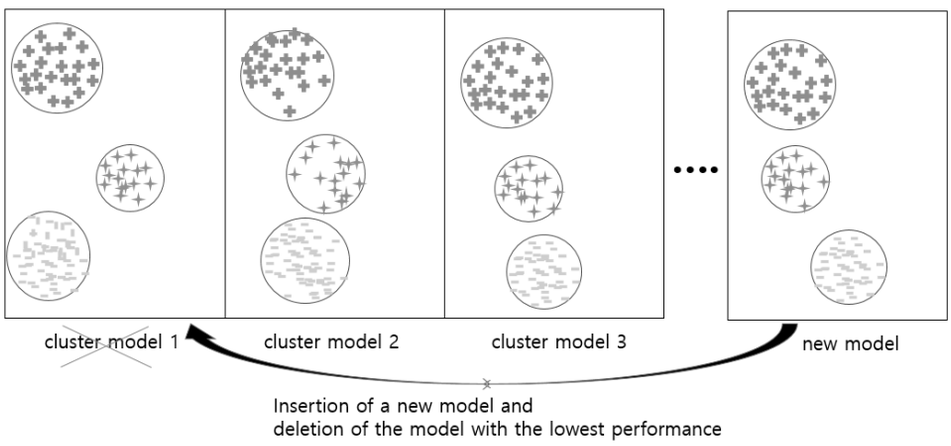


Fig. 3. An Illustration of Clustering-Based Ensemble Method for Outlier Detection on Streaming Data

4. 동적 변화가 있는 스트리밍 데이터에서의 이상치 탐지

모든 데이터가 한꺼번에 주어진 상태에서 학습을 진행하는 배치 모드 학습과 달리 시간에 따라 지속적으로 데이터가 생성되는 스트리밍 모드에서는 데이터를 생성하는 분포가 시간에 따라 변화할 수 있다. 따라서 스트리밍 모드에서는 개념 변동에 따라 학습 모델의 업데이트가 필요하다.

스트리밍 모드에서 앙상블 모델 업데이트를 위해 가장 최근의 데이터 청크로 새로운 모델을 학습하여 앙상블 멤버로 추가하는 방법이 있다. 앙상블 멤버수를 일정하게 유지하고 현재 개념에 맞지 않는 오래된 모델을 제거하기 위해 기존 앙상블 멤버 중에서 가장 성능이 낮은 모델을 앙상블에서 제외한다. 그러나 이를 위해서는 최근에 발생된 데이터 샘플들의 정상/이상치 여부를 어느 정도의 시간 경과 이내에 알 수 있어야 한다는 가정이 필요하다. Fig. 3에서 스트리밍 모드에서의 군집화를 이용한 이상치 탐지 방법을 나타내었다. Fig. 1에서는 초기 앙상블 모델이 변화되지 않는데 반해 Fig. 3에서는 최근 데이터 샘플들로 새로운 모델을 구성하고 기존 앙상블 멤버를 대체하는 것을 보이고 있다.

5. 실험 결과

5.1 실험데이터

이상치 탐지 성능을 평가하기 위한 데이터셀은 실제 이상치를 포함하고 있어야 한다. 그러나 이상치를 포함한 실제 데이터는 얻기는 쉽지 않다. 우리는 다른 논문들에서 많이 사용되는 두 개의 실제 데이터와 두 개의 인공 데이터를 이용하였다. 데이터 샘플의 개수와 이상치 비율 등을 포함한 데이터에 대한 설명을 Table 1에 정리하였다.

Creditcard 데이터는 2013년 9월에 유럽에서 일부 카드 소지자들의 신용 카드 사용 내역을 정리한 것이다[13]. 데이터 샘플의 총 개수는 284,807개이고 그 중 이상치 데이터의 개수는 492개이다. 사용액을 나타내는 속성을 제외하고 28개의 속성을 이용하였다. KDD-http 데이터는 DARPA Intrusion Detection Evaluation Program에 의한 kddcup.data_10_percent_corrected 데이터로부터 얻어졌다. 속성 sevice의 값이 http인 샘플 567,498개에 대해 duration, src_bytes, dst_bytes의 세 개의 속성값을 사용하였다[14]. 네트워크 침입을 나타내는 이상치 데이터는 2211개로 0.39%의 비율이다.

Gaussian 데이터는 1차원 데이터로서 평균이 각각 0, 1, 2, 3, 4인 5개의 가우시안 혼합분포를 이용하여 정상 샘플 25,000개를 생성하고 평균이 6인 가우시안 분포에서 이상치 샘플 2,500개를 추가 생성하였다. RBFevents는 Moa[15]의 인공 스트리밍 데이터 생성자 RandomRBFevents를 이용하여 구성되었다. 5개의 정규분포에서 생성되는 정상 데이터와 랜덤한 균일 분포의 이상치데이터를 포함한다. 시간에 따라 분포변화가 거의 없는 데이터 생성을 위해 speed 변수를 100만으로 하고 speed range를 0으로 설정하여 RBFevents를 생성하였다.

Table 1. The Description of Data Sets

Data	no. of attributes	no. of data samples	no. of outliers	outlier ratio
Creditcard	28	284,807	492	0.17%
KDD	3	567,498	2,211	0.39%
Gaussian	1	27,500	2,500	9%
RBFevents	5	100,000	10,201	10.2%

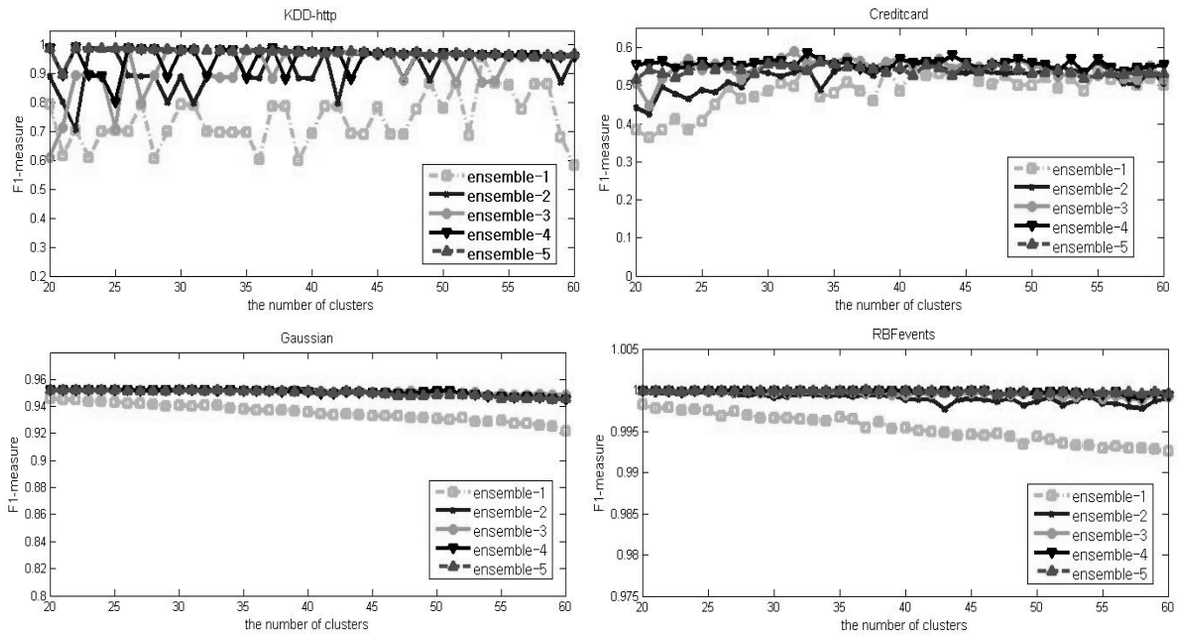


Fig. 4. The Performance Comparison with Respect to the k Value in k-means Clustering and the Number of Ensemble Members

5.2 성능 측정 척도

이상치 탐지 성능은 정확도(precision), 재현률(recall), F1 값으로 평가할 수 있다[16]. 이상치 예측을 positive 판정이라 할 때, 테스트 집합에 있는 데이터 샘플들에 대해 다음과 같이 계산된다.

TP = 이상치를 이상치로 판정한 회수
FP = 정상 데이터를 이상치로 판정한 회수
FN = 이상치를 정상으로 판정한 회수
TN = 정상 데이터를 정상으로 판정한 회수
Precision = TP/(TP+FP)
Recall = TP/(TP+FN)
F1 = 2*Precision*Recall/(Precision + Recall)

Precision이 1에 가까운 값일 때는 이상치로 판정되는 정상 데이터가 거의 없다는 의미이다. 반면에, recall이 1에 가까울수록 이상치를 정상 데이터로 판정하는 FN의 경우가 거의 없다는 것을 나타낸다.

5.3 클러스터 개수와 앙상블 멤버 개수

각 실험 데이터셋에서 30%의 정상 데이터를 이용해 정상 모델을 구성하고 나머지 70%의 데이터로 이상치 예측 성능을 평가하였다. 먼저 클러스터링 모델에서 작은 크기의 클러스터를 삭제하는 것이 노이즈 제거의 효과를 주는가를 테스트하였다. 앙상블 멤버 개수를 3개로 하고 클러스터 개수를 20에서 60까지 변화시켜 가면서 사이즈 5 이하의 클러스터를 삭제할 때와 삭제하지 않을 때의 모델 구성에 의한 이상치 탐지 성능을 비교하였을 때, 5 이하의 사이즈를 가지는 클러스터를 클러스터링 모델에서 제거하는 것이 다양한 실험데이터에서 효과적임을 확인하였다.

Fig. 4에서는 제안 방법에서 클러스터의 개수 k와 앙상블 멤버 개수를 다르게 설정할 때 F1 값을 비교한다. 전반적으로 30~40 이상의 클러스터 개수에서 F1 값이 높아지는 경향을 보였다. 또한, 앙상블 멤버 개수가 많아질수록 클러스터 개수의 영향을 덜 받으면서 성능이 안정되었다.

5.4 다른 방법들과의 성능 비교

기존의 이상치 탐지 방법과의 비교를 위해 이상치 지수를 계산하는 대표적인 방법으로 밀도 기반 방법인 LOF[8], 트리 기반 방법인 Isolation Forest[9]를 비교방법으로 사용하였다. 이상치 여부를 이진으로 판정하는 제안방법과의 비교를 위하여 Table 1에 나타낸 각 실험데이터의 실제 이상치 비율에 해당하는 개수의 이상치 지수 상위에 랭크된 데이터를 이상치로 예측하여 F1 값을 계산하였다. 그러나 실제 환경에서는 이상치 발생 비율을 알 수 없다는 점을 감안하여야 한다. 이상치 여부의 이진판정을 하는 방법과의 비교를 위해서는 one-class SVM을 사용하였다[17]. 정상 데이터를 포함하는 hypersphere를 SVM으로 구성하고 이를 벗어나는 데이터 샘플이 이상치로 판정된다. 모든 비교방법들은 Scikit-learn[18]에서의 구현을 이용하였다.

제안 방법과 one-class SVM, Isolation Forest에서는 각 데이터셋에서 30%의 정상 데이터를 이용해 정상 모델을 구성하고 나머지 70%의 데이터로 성능을 평가하였다. 제안 방법과 one-class SVM은 이상치에 대한 이진판정을 하는 반면, Isolation Forest는 이상치 지수를 계산한다. 따라서 Isolation Forest에서는 70%의 테스트 데이터를 이상치 지수에 따라 정렬하고 실제 이상치 비율을 이용하여 이상치 지수 상위에 있는 데이터를 추출하여 F1 값을 계산하였다. 30%의

Table 2. Performance Comparison by Precision(P), Recall(R), and F1-measure of the Compared Methods

	KDD-http			Creditcard			Gaussian			RBFevents		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Proposed method	**0.969**	0.997	**0.983**	**0.657**	0.492	**0.561**	**0.993**	0.914	**0.952**	**0.9993**	**0.9999**	**0.9996**
one-class SVM	0.896	0.996	0.943	0.122	**0.831**	0.212	0.651	0.93	0.765	0.9927	0.8004	0.8863
LOF	0.053	0.053	0.053	0.02	0.02	0.02	0.1711	0.1712	0.1712	0.1669	0.1669	0.1669
Isolation Forest	0.4875	1	0.6554	0.3727	0.5882	0.4561	0.5257	**0.9752**	0.6831	0.5391	**0.9999**	0.7005

선택을 랜덤하게 10번 반복하여 평균 성능을 계산하였다. LOF는 모든 데이터를 이용해 이상치 지수를 계산하고 실제 이상치 비율을 이용하여 이상치 지수 상위에 있는 데이터를 추출하여 F1 값을 계산하였다.

Table 2에서 비교방법들의 성능을 나타냈다. 제안 방법은 앙상블 개수가 3개이고 k가 30일 때의 값을 표기하였다. one-class SVM은 트레이닝 에러율의 상한을 나타내는 υ값을 0.0001, 0.0005, 0.001, 0.005로 테스트해서 그 중에서 가장 높은 성능일 때의 값을 나타냈다. LOF에서는 이웃의 개수를 5, 25, 50, 100, 200으로 변화시켜 가면서 실험하였고 Isolation Forest는 트리 구성에 필요한 샘플의 수를 나타내는 매개변수를 “auto”로 설정해 min(256, 모든 샘플의 수)을 취하는 경우와 모든 샘플을 사용하는 경우를 실험해서 그 중에서 높은 성능값을 취하였다. 그 외 나머지 매개변수들은 Scikit-learn에서 제시하는 디폴트값을 사용하였다. Table 2에서 보여주는 것처럼 LOF 방법은 실험데이터 대부분의 경우에 낮은 성능을 얻었다. 이웃의 수를 상당히 큰 값으로 했을 때 약간의 성능향상이 있었으나 샘플들 간의 거리 계산으로 인해 오랜 실행시간이 걸리는 경향이 있다. 대부분의 실험데이터에서 제안방법이 다른 방법들과 비슷하거나 높은 성능을 보였다.

제안하는 방법이 클러스터링 모델의 앙상블인데 반해 Isolation Forest는 트리들의 앙상블로 구성된다. 제안하는 방법의 앙상블 멤버 수에 따른 성능비교를 나타낸 Fig. 4에서 보여주듯이 불안정한 모델의 앙상블 구성은 비앙상블 구성보다 더 높은 성능을 줄 수 있다.

5.5 스트리밍 모드에서의 성능 비교

동적변화가 있는 스트리밍 모드에서의 실험을 위해 MOA의 인공 스트리밍 데이터 생성자를 이용하였다. 그러나 Table 1의 RBFevents 데이터의 생성시와 다르게 speed 변수를 5000으로 하고 speed range를 0으로 하여 5개의 정규분포의 중심이 시간에 따라 이동하는 동적 변화를 가진 RBFevents_drift 데이터를 생성하였다. 15만개의 데이터 스트림을 생성하여 처음 3만개의 정상 데이터에 대해 크기 10,000개의 청크마다 클러스터 개수 30으로 하여 k-means clustering을 수행해서 초기 3개의 앙상블 모델을 구성한다. 이후 7만개의 정상과 이상치가 섞여있는 데이터를 이용해 이상치 탐지 성능을 측정한다. 앙상블 모델에 의해 정상/이상치 예측을 실행하면서 정상 데이터 10,000개가 누적되었을때 이를 이용해 새로운 모델을 구성하고 기존 3개의 모델 중에서 예측성능이 가장 낮은 한 개를 대체한다. seed를 변경하면서 데이터 생성을 반복하여 총 10개의 데이터셑을 구성하고 실험을 10번 반복하여 평균 성능을 측정하였다.

Table 3에서 RBFevents_drift 데이터에 대해 앙상블 모델 업데이트 방법과 초기 앙상블 모델을 업데이트하지 않고 그대로 사용하는 방법의 성능을 비교하고 있다. 앙상블 모델을 업데이트하는 경우의 성능이 업데이트 하지 않는 경우에 비해 월등하게 높음을 알 수 있다. 동적 변화가 없는 경우 앙상블 모델 업데이트 방법의 성능은 어떤지 테스트하기 위해 speed 변수를 백만으로 하고 speed range를 0으로 하여 RBFevents_no_drift 데이터 스트림을 생성하여 같은 실험을 수행한 결과가 Table 3에서 보여진다. 데이터 스트림에서 동적변화가 없는 경우 앙상블 모델 업데이트 여부는 성능에 거의 영향이 없음을 알 수 있다.

5.6 윈도우 사이즈의 설정

동적변화가 있는 스트리밍 모드에서 모델구성에 사용되는 데이터 샘플들의 윈도우 사이즈에 따른 성능변화를 실험하기 위해 RBFevents_drift 데이터와 RBFevents_no_drift 데이터를 이용하였다. 처음 3만개의 정상 데이터에 대해 윈도우 사이즈를 1,000으로 설정해서 3개의 클러스터링 모델을 구성하였다. 이후 7만개의 데이터에 대해 앙상블 모델 업데이트 방법과 모델을 업데이트하지 않는 방법의 성능을 비교하였다. 윈도우 사이즈를 1000부터 10,000까지 1,000씩 증가시켜 가면서 실험을 반복하여 Fig. 5에서 비교하였다. RBFevents_no_drift 데이터에서는 윈도우 사이즈가 클수록성능이 높아짐을 볼 수 있다.

Table 3. Performance Comparison of the Ensemble Outlier Detection Method Using RBFevents_drift and RBFevents_no_drift Data

	method	precision	recall	F1-measure
RBFevents_drift	ensemble update	0.8572	0.999	0.9224
	no update	0.1733	0.998	0.2948
RBFevents_no_drift	ensemble update	0.9994	0.9998	0.999
	no update	0.9993	0.9953	0.9973

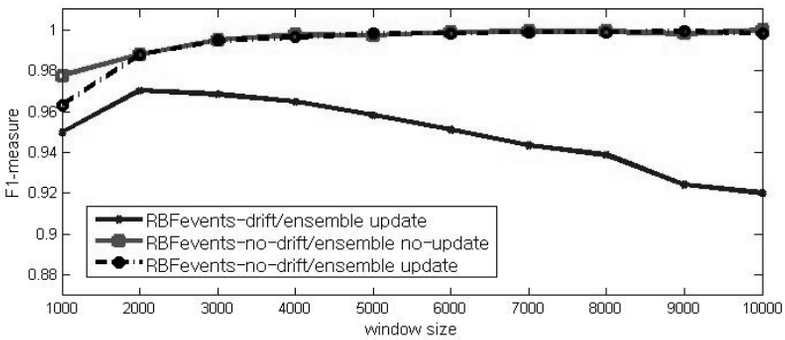


Fig. 5. The Effects of Window Size in the Ensemble Outlier Detection Method Using RBFevents_drift and RBFevents_no_drift data

반면 RBFevents_drift 데이터에서는 윈도우 사이즈가 2000일 때 가장 높은 성능을 보인다. 데이터 특성과 변화속도에 따라 적합한 윈도우 사이즈가 달라질 것이라 예상할 수 있다. 스트림 데이터에 대해서 그 속성을 미리 파악하는 게 쉽지 않을 것이므로 적응적 윈도우 사이즈를 설정할 수 있는 방법의 개발이 필요할 것이다.

6. 결 론

이상치 탐지는 대부분이 정상이라고 여겨지는 데이터집합에서 정상상태를 크게 벗어나는 데이터 샘플을 탐지하는 것을 목적으로 한다. 설비 운용 이상이나 네트워크 침입 탐지 등의 실제 문제에서는 개별적인 데이터 샘플의 정상/이상치 판정을 넘어서 이상 패턴 발생의 실시간 탐지가 필요하다. 그러나 이상패턴 발생 탐지를 위해서는 이상치 예측 성능이 높은 이상치 탐지 방법이 우선적으로 필요하다.

본 논문에서 제안하는 이상치 탐지 방법의 특징을 다음과 같이 요약할 수 있다.

* 정상 데이터만 주어졌을 때 이를 이용하여 정상 데이터 영역을 나타내는 군집화 기반 앙상블 모델을 구성한다.

* 구성된 앙상블 모델을 이용하여 새로운 데이터 샘플에 대해 이상치와 정상치의 이진분류를 수행한다.

또한 동적 변화가 있는 스트리밍 데이터에서도 앙상블 모델 업데이트 방법을 이용하여 높은 이상치 탐지 성능을 가짐을 보였다.

References

[1] E. M. Knorr and R. T. Ng, "Finding intensional knowledge of distance-based outliers," in *Proceedings of 25th International Conference on Very Large Databases*, 1999.

[2] M. Markou and A. Singh, "Novelty detection: a review-part 1: statistical approaches," *Signal Processing*, Vol.83, No.12, pp.2481-2497, 2003.

[3] M. Markou and A. Singh, "Novelty detection: a review - part 2: neural network based approaches," *Signal Processing*, Vol.83, No.12, pp.2499-2521, 2003.

[4] R. Domingues and M. Filippone, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognition*, Vol.74, pp.406-421, 2018.

[5] C. Aggarwal, "Outlier analysis," Springer, Switzerlnd, 2017.

[6] K. Wu, K. Zhang, W. Fan, A. Edwards, and P. Yu, "RS-Forest: A rapid density estimator for streaming anomaly detection," In *Proceedings of the 14th International Conference on Data Mining*, 2014.

[7] F. Angiulli and F. Fassetti, "Detecting distance-based outlies in streams of data," in *Proceedings of CIKM*, 2007.

[8] M. M. Breunig. H-P. Kriegel. R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the 2000 ACM Sigmod International Conference on Management of Data*, 2000.

[9] K. Ting, F. Liu, and Z. Zhou, "Isolation forest," In *Proceedings of the 8th International Conference on Data Mining*, 2008.

[10] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proceedings of the ICML*, 2016.

[11] X. Xu, Z. He, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, Vol.24, pp.1641-1650, 2003.

[12] E. Spinosa, A. Carvalho, and J. Gama, "Olindda: A cluster-based approach for detecting novelty and concept drift in data streams," In *Proceedings of the SAC*, 2007.

[13] Creditcard/Kaggle [Internet], https://www.kaggle.com/arvindratan/creditcard/version/1.

[14] S. Hawkins, H. Hongxing, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery*, 2002.

[15] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis," *Journal of Machine Learning Research*, Vol.11, pp.1601-1604, 2010.

[16] M. Steinbach, P. Tan, and V. Kumar, "Introduction to Data Mining," Addison Wesley, Boston, 2006.

[17] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R.

Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, Vol.13, No.7, pp.1443-1471, 2001.

[18] F. Pedregosa et al, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, Vol.12, pp. 2825-2830, 2011.

박 정 희

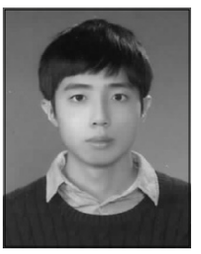
https://orcid.org/0000-0002-8233-2206
e-mail : cheonghee@cnu.ac.kr
1998년 연세대학교 수학과(박사)
2004년 University of Minnesota, Computer Science & Engineering(박사)
2005년～현 재 충남대학교 컴퓨터공학과 교수
관심분야 : 데이터마이닝, 기계 학습, 패턴 인식

김 태 공

https://orcid.org/0000-0002-6720-0408
e-mail : dmflsla@naver.com
2018년 충남대학교 수학과(학사)
2018년～현 재 충남대학교 컴퓨터공학과 석사과정
관심분야 : 데이터마이닝, 기계 학습

김 지 일

https://orcid.org/0000-0003-2688-3887
e-mail : kimjiil2013@naver.com
2018년 충남대학교 컴퓨터공학과(학사)
2018년～현 재 충남대학교 컴퓨터공학과 석사과정
관심분야 : 데이터마이닝, 인공지능



최 세 목

https://orcid.org/0000-0002-4901-3377
e-mail : semok95@daum.net
2018년 충남대학교 컴퓨터공학과(학사)
2018년～현 재 충남대학교 컴퓨터공학과 석사과정
관심분야 : 데이터마이닝, 기계 학습, 자연어 처리

이 경 훈

https://orcid.org/0000-0002-7046-1289
e-mail : ghlee0304@cnu.ac.kr
2017년 충남대학교 수학과(석사)
2016년～현 재 충남대학교 컴퓨터공학과 석사과정
관심분야 : 데이터마이닝, 차원 축소, 딥러닝