Journal of Korea Multimedia Society Vol. 25, No. 6, June 2022(pp. 886-893)
https://doi.org/10.9717/kmms.2022.25.6.886

고차원 데이터에서 One-class SVM과 Spectral Clustering을 이용한 이진 예측 이상치 탐지 방법

박 정 희†

A Binary Prediction Method for Outlier Detection using One-class SVM and Spectral Clustering in High Dimensional Data

Cheong Hee Park†

ABSTRACT

Outlier detection refers to the task of detecting data that deviate significantly from the normal data distribution. Most outlier detection methods compute an outlier score which indicates the degree to which a data sample deviates from normal. However, setting a threshold for an outlier score to determine if a data sample is outlier or normal is not trivial. In this paper, we propose a binary prediction method for outlier detection based on spectral clustering and one-class SVM ensemble. Given training data consisting of normal data samples, a clustering method is performed to find clusters in the training data, and the ensemble of one-class SVM models trained on each cluster finds the boundaries of the normal data. We show how to obtain a threshold for transforming outlier scores computed from the ensemble of one-class SVM models into binary predictive values. Experimental results with high dimensional text data show that the proposed method can be effectively applied to high dimensional data, especially when the normal training data consists of different shapes and densities of clusters.

Key words: Binary Prediction of Outliers, High Dimensional Data, One-class SVM, Outlier Detection, Spectral Clustering

1. 서 론

이상치 탐지는 정상적인 데이터 분포에서 크게 벗어나는 데이터를 탐지하는 작업을 의미한다[1,2]. 이상치 탐지에 대한 연구는 수십년 동안 꾸준하게 이루어져 왔고, 최근에는 네트워크 침입 탐지, 자동 공정 고장 탐지, 불법 카드 사용 탐지, 이상 행동 탐지와 같이 활용범위가 넓어지고 있다[1,2]. 또한, 스트리밍 데이터와 그래프 데이터 등 전형적인 데이터 타입을 벗어나는 다양한 형태의 데이터에 대한 이상치 탐지에 대한 연구도 활발하다[3].

대부분의 이상치 탐지 방법들은 데이터 샘플이 정상에서 벗어나는 정도를 나타내는 이상치 지수(outlier score)를 계산한다[1]. 테스트 데이터 샘플들을 이상치 지수가 높은 것부터 차례대로 정렬하여 상위 높은 랭크에 있는 N개의 데이터 샘플들을 이상치 데이터로 판정하거나 주어진 임계값보다 큰 데이터 샘플을 이상치로 예측한다. 이상치 지수를 계산하는 이상치 탐지 방법은 데이터 샘플이 이상치일 정도를 수치로 나타내 준다는 장점이 있는 반면, 테스트 데이터에 이상치가 포함되어 있는지 알 수 없는 상황일 때 이상치인지 정상 데이터 샘플인지 판단할 수 있는 임계값의 설정이 쉽지 않다는 단점이 있다. 반면에, 이상치 지수를 출력하는 대신에 데이터 샘플이 이상

※ Corresponding Author : Cheong Hee Park, Address: (34134) Daehak-ro 77, Yuseong-gu, Daejeon, Korea, TEL : +82-42-821-6293, FAX : +82-42-822-4997, E-mail : cheonghee@cnu.ac.kr

Receipt date : May 17, 2022, Revision date : Jun. 13, 2022
Approval date : Jun. 16, 2022

† Division of Computer Convergence, Chungnam National University

치인지 정상 데이터인지 나타내는 이진 예측값을 출력하는 이상치 탐지 방법이 있다. 이진 예측값을 주기 때문에 비전문가가 이상치 탐지 결과를 활용하는데 용이할 수 있다.

대표적인 이상치 이진 예측 방법은 클러스터링을 이용하는 방법이다. 클러스터링은 유사도가 높은 데이터 샘플들은 같은 클러스터에 속하게 하고 유사도가 낮은 데이터 샘플들은 서로 다른 클러스터에 속하도록 데이터를 몇 개의 클러스터들로 나누는 작업이다. 유사도는 다양한 거리 척도나 유사도 척도에 의해 계산될 수 있다. 논문[4]에서는 정상 데이터로 구성된 학습데이터가 주어졌을 때 k-means clustering에 의한 클러스터링 앙상블을 구성하고, 테스트 데이터 샘플이 가장 가까운 센터의 클러스터 반경 내에 포함되는지 여부에 따라 정상 데이터인지 이상치인지 예측하는 이상치 이진 예측 방법을 제안하였다. 그러나, 수만 이상의 데이터 차원을 갖는 텍스트 데이터와 같은 고차원 데이터에서는 데이터 샘플들 간의 최대 거리와 최소 거리의 차가 작아지는 차원의 저주(the curse of dimensionality)라고 일컬어지는 현상이 발생하고 거리 기반 클러스터링 방법이 잘 작용하기 어렵다고 알려져 있다[5].

다양한 이상치 탐지 방법들 중에서 one-class SVM을 이용한 이상치 탐지 방법은 분류 경계로부터의 마진을 최대화함으로써 높은 분류정확도를 보이는 SVM(Support vector machine)의 특성에 기반을 두는 방법이다[6]. 그러나, 두 개의 클래스를 나누는 분류경계를 구하는 대신, 주어진 데이터 집합과 원점을 최대의 마진을 가지며 나누는 초평면을 구함으로써 주어진 데이터 영역의 경계를 구한다. 테스트 단계에서는 소수의 서포트 벡터들에 의한 계산을 수행하여 데이터 영역 경계로부터 벗어나는 정도를 가지고 이상치 지수를 계산한다. 데이터가 unimodal 분포를 이루고 있을 때는 one-class SVM을 이용하여 데이터 분포 영역 경계를 구하는 것이 효과적이나, multi-modal 분포를 이루고 있거나 밀도가 다른 영역이 섞여 있는 데이터에 대해서 적용할 때는 커널함수의 파리미터를 조정하기가 쉽지 않다.

본 논문에서는 고차원 데이터에서 그래프 기반 클러스터링 방법인 spectral clustering과 one-class SVM을 이용한 이진 예측 이상치 탐지 방법을 제안한다. 정상 데이터로 구성된 학습데이터가 주어졌을 때, 학습데이터에 spectral clustering을 적용하여 클러스터들을 구하고 각 클러스터에 대해 one-class SVM을 학습하여 앙상블을 구성한다. One-class SVM 모델 앙상블에서 계산된 이상치 지수를 이진 예측 값으로 변환하기 위한 임계값을 구하는 방법을 보인다. 고차원 텍스트 데이터를 이용한 실험을 통하여 제안한 방법이 높은 이상치 이진 예측 성능을 가짐을 입증한다. Fig. 1은 제안 방법의 진행 과정을 보여준다.

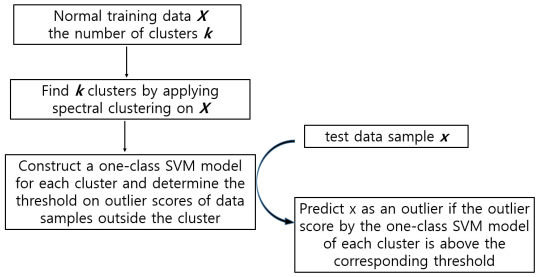

Fig. 1. The flowchart of the proposed method.

본 논문의 구성은 다음과 같다. 2절에서는 이상치 탐지 방법들을 살펴본다. 3절에서는 one-class SVM과 spectral clustering을 이용한 이상치 이진 예측 방법을 제안한다. 4절에서는 고차원 텍스트 데이터를 이용한 실험 결과를 보이고, 5절에서는 결론을 맺는다.

2. 관련연구

정상 데이터와 이상 데이터로 구성된 학습데이터가 주어졌을 때 이상치 탐지 문제는 분류 문제가 된다. 다양한 분류기를 적용하여 정상과 이상치를 예측할 수 있다. 분류 과정에서 예측에 대한 신뢰도를 이용하여 이상치 지수를 계산할 수 있지만, 본질적으로 감독 학습에 의한 이상치 탐지는 이진 예측을 출력으로 한다[1]. 그러나 정상 데이터에 비해서 이상 데이터를 수집하기가 어려운 현실을 감안하면 정상 데이터와 이상 데이터의 양이 크게 차이 나는 불균형 학습의 어려움이 있다.

대부분의 이상치 탐지 방법은 데이터 라벨이 알려지지 않은 데이터셋에 대해서 무감독 이상치 탐지를 수행한다. 가장 잘 알려진 Isolation Forest[7], LOF[8], one-class SVM[6] 등의 이상치 탐지 방법들은 데이터 샘플들이 정상 데이터 분포에서 벗어나 있는 정도를 나타내는 이상치 지수를 계산한다. 거리 기반

이상치 탐지 방법 중에서 대표적인 k-NN 이상치 탐지 방법은 가장 가까운 k개의 이웃들까지의 거리의 합이나 최대 거리를 이용하여 이상치 지수를 계산한다[9]. k개의 이웃과의 거리가 멀수록 이상치일 가능성이 더 커진다. 트리 기반 이상치 탐지 방법들 중에서 가장 잘 알려진 Isolation Forest는 이상치는 정상 데이터에 비해 고립되기 쉽다는 것을 전제로 이진 트리들의 앙상블을 구성한다. 주어진 데이터에서 랜덤 샘플링에 의한 부분 집합을 가지고 임의로 선택되는 속성에 의한 이진 분할을 모든 데이터 샘플이 분리되거나 트리가 정해진 높이에 도달할 때까지 반복하여 이진 트리를 구성한다. 각 데이터 샘플의 이상치 지수는 루트 노드로부터 리프노드까지 거쳐가는 경로 길이를 기반으로 계산된다. 경로 길이가 짧을수록 고립되기 쉬운 이상치일 가능성이 크게 된다. 이상치 지수를 출력값으로 가지는 탐지 방법에서는 가장 큰 이상치 지수를 가진 N개의 데이터 샘플을 구하거나 또는 임계값을 설정해 임계값보다 큰 이상치 지수를 가지는 데이터 샘플들을 이상치로 판단할 수 있다. 그러나 일반적으로 이상치 지수가 정규화되어 있지 않고 임계값의 설정에 대한 신뢰성 있는 방법이 없어서 이상치 지수로부터 이상치 이진 예측을 결정하기는 어렵다. 주어진 데이터셋에 대해 클러스터링을 수행하여 다른 클러스터들에 비해 작은 사이즈의 클러스터에 속하는 데이터 샘플들을 이상치라고 판단하기도 한다.

정상 데이터만으로 구성된 학습데이터가 주어질 때, 테스트 데이터에 대해 정상인지 이상치인지 예측하는 방법으로, 논문[4]에서는 클러스터링 앙상블에 기반한 이상치 이진 예측 방법을 제안하였다. 정상 학습데이터를 몇 개의 청크로 나누고, 각 청크에 대해 k-means clustering을 수행하여 정상 데이터 영역을 하이퍼스피어(hypersphere)들의 합집합으로 표현한다. 테스트 데이터 샘플이 주어졌을 때, 각 청크의 k-means clusteing에 의한 클러스터들 중에서 가장 가까운 클러스터 센터의 하이퍼스피어 반경에 포함되지 않을 때 이상치로 판단되고, 모든 청크에서 이상치로 예측될 때 최종적으로 이상치로 결정한다. 논문[10]에서는 k-means clustering에서처럼 가장 가까운 클러스터 센터에 할당되는 데이터 샘플들로 새로운 센터를 구하는 과정을 반복하면서 클러스터리을 수행한다. 그러나 하이퍼스피어 대신에 SVDD (Support vector domain description)[11]에 의한 클러스터 경계를 구한다. 테스트 데이터 샘플에 대해 가장 가까운 센터를 가지는 클러스터의 경계 안에 들어갈 때 정상 데이터로 판정한다.

3. One-class SVM과 Spectral Clustering을 이용한 이진 예측 이상치 탐지 방법

k-means clustering은 샘플들과 가장 가까운 클러스터 센터와의 거리를 최소화하도록 클러스터를 구성하는 클러스터링 방법이다. 그러나 고차원 데이터에서는 데이터 쌍의 거리들이 크게 차이 나지 않는 경향이 있어 k-means clustering의 적용이 데이터에서의 클러스터 구조를 얻는데 어려울 수 있다. 반면에, spectral clustering은 이웃 간의 연결을 에지로 나타내는 그래프 기반 클러스터링 방법으로 non-convex 형태의 클러스터들이 존재할 때 효과적으로 클러스터링을 수행할 수 있다[12,13]. k-means clustering으로 생성되는 클러스터는 센터와 반경의 크기로 클러스터를 특징지을 수 있으나 spectral clustering으로 생성되는 클러스터는 다양한 형태를 가질 수 있다. 이러한 문제를 해결하고자 one-class SVM을 적용하여 클러스터 영역의 경계를 묘사하고자 한다.

고차원 데이터에서 정상 데이터로 구성된 학습데이터가 주어졌을 때 이진 예측 이상치 탐지를 위해, 먼저 spectral clustering을 적용하여 클러스터들을 구한다. 이후 각 클러스터에 대해 one-class SVM을 적용하고 이상치 이진 예측을 위한 이상치 지수에 대한 임계값을 구한다. 각 클러스터에서 구성된 one-class SVM 모델의 앙상블에서 모두 이상치로 예측될 때 최종적으로 이상치로 판단한다. 다음에서 제안 방법의 과정들을 자세히 설명한다.

3.1 Spectral clustering을 이용한 정상 학습데이터 클러스터링

Spectral clustering은 데이터 샘플과 이웃들을 에지로 연결한 그래프를 구성한다. 그래프를 표현하는 유사도행렬의 고유벡터들에 의한 임베딩 공간에서 k-means clustering과 같은 알고리즘을 적용해서 클러스터링을 수행한다[12,13]. 클러스터의 형태가 컨벡스(convex)형태가 아닐 때 유용하게 적용될 수 있는 클러스터링 방법으로 알려져 있다. Spectral clus-

tering의 과정을 정리하면 다음과 같다.

1. 유사도 행렬(affinity matrix) $W=[w(i,j)]$를 구성한다.
   - 각 데이터 샘플마다 s개의 가까운 이웃들과의 에지들을 구성해서 행렬 W를 구성한다.
   - $W \leftarrow 0.5*(W+W^T)$
2. 라플라시안 행렬 $L=D-W$에 대해서 일반화된 고유값 문제(generalized eigenvalue problem) $Lu=\lambda Du$에서 0을 제외한 가장 작은 K개의 고유값에 해당하는 고유벡터 $u_1, u_2, \ldots, u_K$를 칼럼으로 가지는 행렬 U를 구한다. 여기서 D는 $d_i=\sum_j w(i,j)$를 원소로 가지는 대각행렬이다.
3. 행렬 U의 각 행이 데이터 샘플이 K차원으로 임베딩된 벡터에 해당된다. U의 행벡터들에 대해 k-means clustering을 수행한다.

첫 번째 스텝인 유사도 행렬을 구성할 때 가우시안 커널을 이용하여 가중치 유사도 행렬을 구성할 수도 있다. 본 논문의 실험에서는 Scikit-learn[14]을 사용하여 spectral clustering을 구현하였고, 디폴트 값인 s=10을 사용하여 이웃과의 연결 구조를 이용하는 무가중치 유사도 행렬을 구성하였다.

3.2 One-class SVM 앙상블 구성

One-class SVM은 주어진 데이터 $\{x_1, \ldots, x_n\}$을 커널함수에 대응하는 공간으로 매핑하고 원점으로부터 가장 큰 마진을 가지고 나누는 경계를 구하기 위해 식(1)의 최적화 문제를 푼다.

$$\min_{\{w,\rho,\xi\}} \frac{1}{2}\|w\|^2+\frac{1}{\nu n}\sum_i \xi_i-\rho \quad subject\ to\ w\cdot\Phi(x_i)\geq\rho-\xi_i,\ \xi_i\geq 0. \tag{1}$$

ξ_i는 경계 $w\cdot\Phi(x)-\rho=0$의 음의 영역에 있는 데이터 샘플들에 대한 패널티로 작용하는 슬랙 변수이고, ν는 데이터 영역 경계를 벗어나게 되는 데이터 샘플들의 비의 상한을 의미한다. 함수 Φ에 의해 매핑된 공간에서 $w=\sum_j w_j\Phi(x_j)$로 표현되고 내적은 커널 함수 $\kappa(x_i, x_j)=\Phi(x_i)\cdot\Phi(x_j)$를 통하여 계산되므로 $w\cdot\Phi(x)=\sum_j w_j\kappa(x_j,x)$로 구하게 된다. One-clas SVM 모델에 의한 이상치 지수는 $-1*(w\cdot\Phi(x)-\rho)$의 값으로 계산되며 데이터 영역 경계로부터 음의 영역으로 멀리 떨어져 있을수록 더 큰 값을 가지게 된다.

Spectral clustering을 적용하여 얻은 클러스터 각각에 대해 one-class SVM을 적용한다. 4절의 실험에서는 Scikit-learn의 OneClassSvm() 함수를 이용하여 구현하였고, 모든 파라미터는 디폴트 값을 사용하였다. 예를 들어, rbf 커널함수 $\exp(-\gamma\|x_1-x_2\|^2)$의 γ는 1/(속성 개수*클러스터에 속하는 데이터의 variance)로 설정되었다.

3.3 이진 예측 이상치 탐지를 위한 임계값 결정

이제 one-class SVM 앙상블을 이용하여 테스트 데이터가 이상치인지 정상 데이터인지 예측하기 위한 임계값을 설정해야 한다. i번째 클러스터에 대해 모델링된 one-class SVM에 의한 데이터 샘플 x의 이상치 지수 $f_i(x)$는 그 클러스터 영역으로부터 벗어나는 정도를 나타내는 값이다. 클러스터 영역 내에 있는 데이터 샘플들과 영역 밖에 놓이는 데이터 샘플들을 나누는 f_i값에 대한 임계값 s_i를 설정하기 위해 다음과 같은 과정을 적용하였다.

1. 클러스터 i에 속하는 데이터 샘플 중에서 랜덤하게 90%를 선택하여 데이터셋 $train_i$를 구성하고 나머지 10%를 데이터셋 val_i에 둔다.
2. 클러스터 i에 속하지 않는 데이터 샘플들 중에서 랜덤하게 10%를 선택해서 val_i에 추가한다.
3. $train_i$에 있는 샘플들의 f_i 값들의 p-번째 백분위수를 임계값으로 하여 val_i에 있는 데이터 샘플들의 f_i 값에 적용했을 때 클러스터 i에 속하는 데이터와 속하지 않는 데이터의 예측성능이 가장 높은 p를 구한다. 실험에서는 80, 85, 90, 95, 100번째 백분위수 중에서 val_i에 대한 예측성능을 가장 높게 주는 p값을 구하고 p-번째 백분위수를 임계값 s_i로 설정하였다.

테스트 샘플이 이상치로 예측되기 위해서는 모든 클러스터 영역 밖에 있어야 한다. 즉 모든 i에 대해 $f_i(x)>s_i$이면 테스트 데이터 샘플 x는 이상치로 예측한다.

4. 실험결과

4.1 실험 데이터

제안한 이진 예측 이상치 탐지 방법의 성능 비교

를 위하여 텍스트 데이터셋을 이용하였다. Table 1은 사용된 데이터에 대한 설명을 보여준다. BBC 뉴스 데이터는 BBC에서 작성된 2,225개의 뉴스 데이터로 다섯 개의 카테고리 business, entertainment, politics, sport, tech로 구성되어 있다[15]. 논문[2]에서와 같은 데이터 전처리 과정을 적용하여 특수기호와 숫자를 삭제하고 한 개의 문서에만 속하는 단어를 제거하여 17,005개의 용어를 가진다. Reuters 데이터는 UCI machine learning repository에서 다운받은 Reuters-21578 데이터에서 135개의 TOPICS 카테고리에 속하는 다큐먼트들을 사용하였다[2]. Stopwords removal, stemming, tf-idf transformation, unit norm을 가지도록 전처리하고, 두 개 이상의 그룹에 속하는 문서를 제외하여, 15,484개의 용어로 구성되는 6,656개의 다큐먼트를 가지게 된다[2]. 135개의 카테고리 중에서 가장 많은 다큐먼트들을 가지는 1과 36 카테고리와 나머지 모든 문서를 모은 카테고리로 총 세 개의 클래스를 구성하였다[2]. 20-newsgroup 데이터는 20개의 뉴스 그룹과 5개의 카테고리로 나누어진 20000개의 기사를 포함하는 20news-bydate version을 사용하였다[16]. Reuters 데이터에서와 같은 전처리를 통해 44,713개의 용어로 구성되는 18,774개의 다큐먼트로 구성되었다. Medline은 의학 논문의 초록들을 모은 데이터베이스에서 추출된 텍스트 데이터를 전처리한 데이터이다[17]. 나머지 5개의 데이터셋은 공공으로 사용가능한 사이트에서 다운받아서, 각 데이터셋에서 200개보다 적은 데이터를 가지는 클래스와 한 번 이하의 빈도수를 가지는 용어를 제거하여 구성되었다[18].

각 데이터셋에서 한 개의 클래스를 이상치 클래스로 두고 나머지 클래스를 정상 데이터로 설정하는 멀티 클래스 정상 데이터 환경과 두 개의 클래스를 선택해 하나는 이상치 클래스로 다른 하나는 정상 클래스로 설정하는 원 클래스 정상 데이터 환경의 두 가지 환경에서 실험을 수행하였다. 실험 성능은 테스트 데이터에서 이상치 예측에 대한 정확도, 재현률과 f1 값으로 측정하였다.

Table 1. The description of data.

data	dim	samples	classes
BBC	17005	2225	5
Reuters	15484	6656	3
20newsgroup	44713	18774	20
medline	22095	2500	5
la12	21604	6279	6
sports	18324	8313	5
classic	12009	7094	4
ohscal	11465	11162	10
reviews	23220	3932	4

정확도(precision): 이상치로 예측된 데이터 중에서 실제 이상치인 데이터 비율

재현률(recall): 실제 이상치 중에서 이상치로 예측이 된 데이터의 비율

f1: 정확도와 재현률의 조화평균, 즉 (precision + recall)/(2*precision*recall)로 계산.

4.2 멀티 클래스 정상 데이터 환경에서의 이진 예측 이상 탐지 성능 비교

멀티 클래스로 구성된 정상 데이터 환경에서의 이상 탐지 성능 비교를 위해 각 데이터셋에서 한 개의 클래스를 이상치 클래스로 두고 나머지 클래스를 정상 데이터로 설정하였다. 정상 데이터의 각 클래스에서 50%의 데이터 샘플을 랜덤하게 선택해서 학습 데이터를 구성하고, 나머지 50%의 데이터들과 이상치 클래스에서 50%의 데이터를 선택해서 테스트셋을 구성하였다. 정상 학습데이터로부터 이상치 탐지 모델을 학습하고, 테스트 데이터에 대해 이상치 예측을 수행하여 f1 값을 구할 수 있다. 어떤 클래스를 이상치 클래스로 설정하느냐에 따라 성능이 달라질 수 있으므로 각 데이터셋에서 이상치 클래스로 설정하는 클래스를 다르게 설정하면서 20번 실험을 반복 수행하여 평균 f1 값을 측정하였다.

이진 예측 이상치 탐지 성능 비교를 위해 다음 방법들과 성능을 비교하여, Table 2에서 실험결과를 보여준다.

1. k-means clustering ensemble[4]: 논문 [4]에서 설정한 것과 같이 앙상블의 멤버 개수는 3으로 하였다. k-means clustering에서 클러스터의 개수는 20, 30, 40으로 다르게 수행한 후 가장 성능이 높은 값으로 Table 2에 나타냈다.

2. One-class SVM: 학습데이터 전체에 대해 one-class SVM을 적용하여 모델링한 후, 다음 두 가지 방법에 의해 이상치에 대한 이진 예측을 수행하였다.

Table 2. Performance comparison by f1 values in the outlier detection simulation of multi-class normal training data.

data	k-means clustering ensemble [4]	One-class SVM		k-means clustering+ one-class SVM ensemble		spectral clustering+one-class SVM ensemble	
		Zero	EST	Zero	EST	Zero	EST
BBC	0.524	0.394	0.407	0.359	0.569	0.364	**0.661**
Reuters	0.501	0.686	0.691	0.669	0.696	0.675	**0.73**
20ng	0.245	0.437	**0.438**	0.434	0.307	0.435	0.431
medline	0.389	0.361	0.364	0.339	0.411	0.34	**0.464**
la12	**0.478**	0.35	0.352	0.325	0.418	0.337	0.41
sports	0.362	0.468	0.47	0.44	0.484	0.451	**0.623**
classic	**0.671**	0.477	0.486	0.447	0.609	0.455	0.648
ohscal	0.194	0.22	0.221	0.217	0.232	0.219	**0.238**
reviews	**0.519**	0.473	0.477	0.439	**0.519**	0.46	0.5
average	0.432	0.43	0.434	0.407	0.472	0.415	**0.523**

[Zero(zero threshold)] 테스트 데이터에 대해 one-class SVM에 의한 이상치 지수가 0보다 클 때 이상치로 예측한다.

[EST(estimated threshold)] 학습데이터의 이상치 지수 값들에 대해 p번째 백분위수로 임계값을 정하고 테스트 데이터에 대한 이상치 지수가 임계값보다 클 때 이상치로 예측한다. 80, 85, 90, 95, 100번째 백분위수를 가지고 실험한 후 테스트 데이터에 대한 가장 높은 f1 값을 Table 2에 나타내었다.

3. k-means clustering + one-class SVM ensemble: k-means clustering을 이용한 정상 학습데이터 클러스터링을 수행한 후 one-class SVM 앙상블에 의한 이진 예측 이상치 탐지를 위한 임계값 결정을 다음 두 가지 방법에 의해 수행한다.

[Zero(zero threshold)] 테스트 데이터에 대해 one-class SVM 앙상블의 각 멤버에 의한 이상치 지수가 0보다 클 때 이상치로 예측한다.

[EST(estimated threshold)] 3.3절의 방법에 의해 이진 예측 이상치 탐지를 위한 임계값을 결정한다.

4. Spectral clustering + one-class SVM ensemble: spectral clustering을 이용한 정상 학습데이터 클러스터링을 수행한 후 one-class SVM 앙상블에 의한 이진 예측 이상치 탐지를 위한 임계값 결정을 3번 방법에서와 같이 [Zero(zero threshold)]와 [EST (estimated threshold)]에 의해 각각 수행한다.

모든 비교 방법들은 Scikit-learn[14]과 Python 3.7을 이용하여 구현하였고, CPU 3.5GHz, RAM 32 GB, Windows 10 환경에서 수행하였다. 정상 학습데이터 클러스터링에서 클러스터의 개수를 결정하기 위해, BBC 데이터를 이용한 사전실험에서 클러스터의 개수를 2에서 9까지 변화시켜 가면서 실험을 수행하여 비교한 후, 클러스터의 개수를 6으로 고정하여 전체 데이터셋에서 적용하였다. 학습데이터가 정상 데이터로 구성되어 있으므로 one-class SVM에서 ν 값을 0.01로 하였고 그 외의 다른 모든 파라미터 값은 Scikit-learn에서의 디폴트 값을 이용하였다.

Fig. 2에서 비교 방법들의 평균 f1 vlaue를 나타내고 있다. Fig. 2에서 보여주는 것처럼 spectral clustering과 one-class SVM 앙상블에 의한 방법은 전반적으로 다른 방법들과 비교하여 높은 성능을 보여주었다. K-means clustering이나 one-class SVM을 단독으로 사용하는 방법들에 비해서 평균 20% 높은 f1 value를 얻었다. 또한, 이상치 지수에 대한 임계값을 결정할 때 단순히 0을 기준으로 정하는 것보다 정상 학습 데이터의 클러스터링을 이용한 임계값 예측이 k-means clustering+one-class SVM ensemble 방법에서는 평균 15%의 f1 value를 향상시켰고, Spectral clustering+one-class SVM ensemble 방법에서는 평균 26% 높은 f1 value를 얻었다. 반면에 k-means clustering과 one-class SVM 앙상블에 의한 방법은 논문 [4]에서의 방법보다는 높으나 spectral clustering을 사용했을 때보다는 낮은 성능을 보였다. 이러한 결과는 고차원 데이터에서 이웃간의 연결을 이용한 그래프 기반 spectral clustering이 클러스터의 구조를 좀 더 잘 파악함을 의미하고, one-class SVM 앙상블에 의한 이진 예측 방법이 단순히 클러

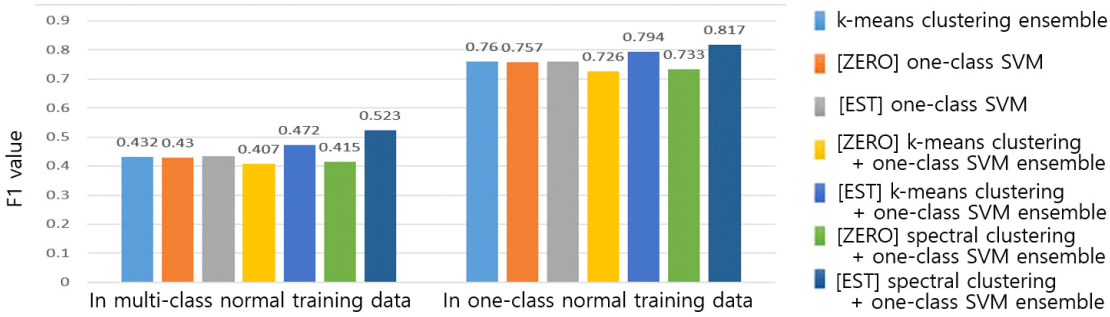

Fig. 2. The comparison of average f1 values of the compared methods.

Table 3. Performance comparison by f1 values in the outlier detection simulation of one-class normal training data.

data	k-means clustering ensemble [4]	One-class SVM		k-means clustering+one-class SVM ensemble		spectral clustering+one-class SVM ensemble	
		Zero	EST	Zero	EST	Zero	EST
BBC	0.78	0.738	0.753	0.7	0.867	0.704	**0.875**
Reuters	0.684	0.846	0.848	0.83	**0.856**	0.837	0.845
20ng	0.604	0.743	**0.745**	0.739	0.538	0.743	0.655
medline	0.7	0.69	0.693	0.669	0.723	0.671	**0.795**
la12	0.827	0.736	0.74	0.695	0.823	0.717	**0.832**
sports	0.772	0.792	0.794	0.747	**0.872**	0.752	0.851
classic	0.835	0.763	0.771	0.721	**0.876**	0.729	0.856
ohscal	0.816	0.759	0.762	0.719	0.818	0.726	**0.834**
reviews	**0.826**	0.742	0.746	0.716	0.775	0.72	0.814
average	0.760	0.757	0.761	0.726	0.794	0.733	**0.817**

스터 반경을 이용하는 경우보다 정밀하게 클러스터 영역을 나타낼 수 있음을 보여준다.

4.3 원 클래스 정상 데이터 실험 환경에서의 이상 탐지 성능 비교

멀티 클래스 정상 데이터 환경에서의 이상 탐지와 달리 이번 실험에서는 한 개의 클래스를 정상 클래스로 두고 다른 한 개의 클래스를 이상치 클래스로 설정하였을 때 이상 탐지 성능을 비교한다. 정상 클래스로 설정된 클래스에서 랜덤하게 90%의 데이터를 선택하여 정상 학습데이터셋을 구성하고 나머지 데이터와 이상 클래스에서 10% 선택된 데이터를 합하여 테스트 셋을 구성하였다. 모든 클래스의 쌍이 정상과 이상 클래스로 역할을 하도록 반복해서 실험을 하여 평균 f1 값을 측정하였다.

Table 3와 Figure 2에 요약된 실험 결과는 클러스터링과 one-class SVM 앙상블을 기반으로 한 제안 방법이 k-means clustering ensemble이나 one-class SVM을 단독으로 사용하는 방법들에 비해 평균 0.03 ~0.05 높은 f1 value를 얻었다. 그러나 멀티 클래스 정상 데이터 환경에서와 달리 원 클래스 정상 데이터 경우에는 k-means clustering을 사용한 one-class SVM 앙상블의 성능이 spectral clustering을 사용했을 때와 차이가 적어졌음을 볼 수 있다.

5. 결 론

본 논문에서는 spectral clustering과 one-class SVM 앙상블을 이용한 이진 예측 이상치 탐지 방법을 제안하고, 텍스트 데이터를 이용한 실험을 통해 이상치 탐지 성능이 향상됨을 보였다. 특히 멀티 클래스로 구성된 정상 데이터 환경에서 k-means clustering을 이용하는 방법들에 비해 평균 20% 높은 f1 value를 얻었다. 이는 고차원 데이터에서 클러스터

의 형태나 데이터 밀도가 다른 영역들이 혼재할 때 spectral clustering으로 클러스터 구조를 파악하고 이를 바탕으로 one-class SVM을 적용한 클러스터 데이터 영역 표현이 효과적으로 이루어질 수 있음을 입증한다. 또한, 정상 데이터의 클러스터링을 이용하여 one-class SVM 모델의 이상치 지수에 대한 임계값을 설정하는 방법이 이상치 지수 출력을 이진 예측 이상치 탐지로 변환하는데 효과적임을 보여준다.

REFERENCE

[1] C. Aggarwal, *Outlier Analysis*, Springer, Switzerland, 2017.

[2] C. Park, "A Distance-based Outlier Detection Method Using Landmarks in High Dimensional Data," *Journal of Korea Multimedia Society*, Vol. 24, pp. 1242-1250, 2021.

[3] C. Park, "Outlier and Anomaly Pattern Detection on Data Streams," *The Journal of Supercomputing*, Vol. 75, pp. 6118-6128, 2019.

[4] C. Park, T. Kim, J. Kim, S. Choi, and G. Lee, "Outlier Detection by Clustering-based Ensemble Model Construction," *KIPS Transactions on Software and Data Engineering*, Vol. 7, pp. 435-442, 2018.

[5] N. Singh, N. Garg, and J. Pant, "A Comprehensive Study of Challenges and Approaches for Clustering High Dimensional Data," *International Journal of Computer Applications*, Vol. 92, 2014.

[6] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the Support of a High-dimensional Distribution," *Neural Computation*, Vol. 13, pp. 1443-1471, 2001.

[7] F. Liu, K. Ting, and Z. Zhou, "Isolation Forest," *Proceeding of ICDM*, 2008.

[8] M. Breunig, H. Kriegel, R. Ng, and J. Sander, "LOF: Identifying Density-based Local Outliers," *Proceeding of the ACM Sigmod International Conference on Management of Data*, 2000.

[9] S. Damaswanny, R. Rastogi, and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," *Proceeding of ACM SIGMOD*, pp. 427-438, 2000.

[10] S. Lyu and H. Farid, "Steganalysis Using Color Wavelet Statistics and One-class Support Vector Machines," *Proceeding of IS&T/SPIE Electronic Imaging*, 2004.

[11] D. Tax and R. Duin, "Data Domain Description Using Support Vectors," *Proceeding of European Symposium on Artificial Neural Networks*, 1999.

[12] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 888-905, 2000.

[13] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, Vol. 15, pp. 1373-1396, 2003.

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

[15] D. Greene and P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," *Proceeding of ICML*, 2006.

[16] 20 Newsgroups(2008), http://qwone.com/~jason/20Newsgroups/ (accessed June 27, 2022).

[17] H. Kim, P. Holland, and H. Park, "Dimension Reduction in Text Classification with Support Vector Machines," *Journal of Machine Learning Research*, Vol. 6, pp. 37-53, 2005.

[18] Karypis Lab, http://glaros.dtc.umn.edu/gkhome/index.php (accessed June 27, 2022).

박 정 희

1998년 연세대학교 수학과 (박사)
2004년 University of Minnesota, Computer Science & Engineering(박사)
2005년~현재 충남대학교 컴퓨터공학과 교수

관심분야 : 기계학습, 데이터마이닝