클래스 분포 변화 탐지[1)]

심홍석 박정희
충남대학교 컴퓨터공학과
shim1320@cnu.ac.kr, cheonghee@cnu.ac.kr

Detection of An Evolving class

Hongsuk Shim Cheonghee Park
Dept. of Computer Science and Engineering
Chungnam National university

감독학습(supervised learning)을 기반으로 한 분류화 과정에서 클래스 분포가 변화했을 때 새로운 경향을 반영하여 분류화 모델을 수정할 수 있어야 한다. 새로운 데이터가 기존의 클러스터 구조에 적합한지 판단하는 것은 다양한 방법으로 연구되어 왔다. 감독학습(supervised learning)에서는 분류기에 거부 옵션 (reject option)을 주어서 기존의 클러스터 구조와의 적합성을 판단할 수 있다[1]. 이에 반해 무감독학습(unsupervised learning)에서의 아웃라이어(outlier)탐지는 기존의 데이터 분포로부터 비정상적으로 벗어나 있다고 생각되는 데이터를 찾는 것이다[2]. 거부 옵션이나 아웃라이어 탐지는 개별의 데이터샘플에 대한 처리이므로, 기존의 분류모델에 맞지 않는 데이터들을 분류할 수는 있지만 탐지된 데이터들로부터 새로운 패턴을 가진 클래스의 발생을 탐지하기는 어렵다. 일반적인 분류기법들이 학습데이터에 포함되지 않았던 새로운 패턴의 데이터를 기존의 데이터와 구별할 수 있는 성능을 가지고 있지 않으며, 데이터 내에 노이즈가 있을 수 있기 때문에 단순히 분류가 거부되어진 데이터가 새로운 클래스를 형성하는 데이터라고 가정하는 것은 좋지 않은 방법이다. 이 논문에서는 데이터 분포의 변화를 감시하고 새로운 클래스 생성을 탐지하는 시스템을 소개한다. 먼저 주어진 클러스터구조에 맞지 않는 데이터들의 발생빈도가 커질 때 데이터분포의 변화를 알려줄 수 있는 통계적 가설 검증(statistical hypothesis testing)방법을 제안한다. 가설 검정에 의해 새로운 클래스 발생에 대한 경고가 울리면 새로운 클래스의 원소라고 생각되는 후보자들을 찾게 된다.

다양한 분류(classification)기법중에서 어떠한 클래스에 속하는지에 대한 결정대신 각각의 클래스에 속할 신뢰도(confidence measure)를 측정하기도 한다[1,3]. 최근에 제안된 분류법으로서 TCM-kNN (Transduction confidence machine for k nearest neighbors)은 클래스에 대한 예측과 함께 클래스 예측의 신뢰도를 측정하기 위해 transduction에 바탕을 둔 kNN 분류기를 사용한다[3]. TCM기반의 아웃라이어 탐지방법인 StrOUD[4]는 클러스터 알고리즘을 사용하여 클러스터링을 수행한 후 형성된 클러스터 구조에 대해서 아웃라이어인지를 판별한다. 변화하는 데이터에서 기존의 클러스터 구조에서 벗어나는 새로운 타입의 데이터 샘플이 발생하면, 이러한 데이터들의 각 클래스에 대한 신뢰도(confidence)는 클래스별로 잘 분류된 다른 데이터 샘플보다 낮을 것으로 기대할 수 있다. 그러나 기존의 클래스에 속하는 데이터 샘플들이 아웃라이어처럼 작은 값을 가질 수도 있으므로 기존의 클러스터의 변화를 경고하기 위해 단순히 임계치보다 낮은 데이터 샘플들의 발생회수를 새는 것은 적합하지 않다. 이러한 문제점을 해결하기 위해서 우리는 통계적 가설 검증시스템을 구축한다. 먼저, 기존의 클러스터 분포에서 발생한 데이터가 낮은 신뢰도를 가짐으로써 아웃라이어로 분류될 확률 q_0를 TCM을 기반으로 하여 설정한다. 그리고 테스트 샘플들의 분류과정에서 아웃라이어로 판정되는 발생비율 q가 클래스 분포변화에 대해 통계적으로 의미가 있는지를 판단하기 위해 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)을 $H_0: q = q_0$, $H_1: q > q_0$과 같이 세울수 있다. 테스트 샘플들의 숫자 N이 충분히 클때 q는 정규분포 $N(q_0, \frac{q_0(1-q_0)}{N})$를 따른다[5]. 따라서 검정통계량(test statistic)은 $z = \frac{q - q_o}{\sqrt{\frac{q_0(1-q_0)}{N}}}$ 과 같이 표현 될 수 있다. 유의 수준 α에 대해서, 만약 $z > z_\alpha$ 라면 귀무가설은 기각

1) 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국 학술 진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-331-D00510)

되고 기존 클러스터 분포의 변화에 대한 경고가 울리게 된다. 위의 통계적 가설 검정은 단측 검정이므로 단측 검정에 대한 z_a값들을 사용한다. 우리는 데이터 분포 변화 감지를 위한 제안된 알고리즘을 HT-TCM이라 부른다. 위 실험을 위해서 UCI[6]로부터 다운 받은 Letter recognition data(Imgchar)와 Isolet recognition data(Isolet)를 사용해서 HT-TCM-OUD의 성능을 평가했다. HT-TCM-OUD는 StrOUD에서 사용했던 신뢰도(confidence measure)를 HT-TCM에 적용한 것이다. Imgchar에 대해서 번갈아 가면서 26개의 클래스들 중 한 개의 클래스를 새로이 생성되는 클래스로 설정하고, 각각의 경우 트레이닝 셋을 25개의 각 클래스로부터 4/5를 랜덤하게 추출해서 구성하고 나머지 1/5를 테스트 셋으로 사용했다. 또한 새로운 클래스로 설정된 클래스 중 1/5를 테스트 셋에 추가시켰다. q_0=0.05로 하고 유의 수준 a = 0.05 즉, z_a = 1.64로 하여 테스트를 하였다. 그 결과 26개의 경우중에 5가지 경우를 제외하고는 새로운 클래스로 인한 분포 변화를 탐지했다. 유의수준 a = 0.1로 하면 25개의 경우에 대한 경고를 주었다. 또한 새로운 클래스가 포함되지 않은 경우에는 어떤 경우에도 경고를 울리지 않았다. Isolet에 대한 실험에서는 21개의 경우에 대해서 새로운 클래스 분포를 탐지 하지 못했다. 그 이유로서 서로 다른 클래스 간의 겹침(overlapped data)으로 인해 새로운 클래스 탐지를 어렵게 만들었다고 추측한다.

클러스터구조에서 변화를 감지하는 신호가 울리면 새로운 클래스에 속할 가능성이 있는 테스트 샘플들을 탐지하여 시스템 전문가등의 면밀한 검토를 요구할 필요가 있다. 새로운 특징을 갖는 클래스가 출현하게 되면 새로운 클래스로부터 나온 원소는 랜덤하게 퍼져있기보다는 어느 정도 일정한 밀도를 갖는 클러스터를 구성할 것이다. HT-TCM-OUD방법을 통해 아웃라이어로 판정된 테스트 샘플들 중에서 밀도가 높은 클러스터를 감지하기 위해서 우리는 격자분할(grid cells)을 사용했다. 데이터 차원공간을 격자로 표현하고 각 단위 공간 안에 속하는 원소의 개수를 센다. 그러나 이 방법의 문제점은 고차원의 공간이나 희박한 데이터분포에서는 밀도가 높은 구역을 찾기는 힘들다는 것이다. 또한 고차원공간에서는 격자 방법은 시간과 메모리차원에서 매우 비용이 많이 든다. 좀 더 빠른 시간과 작은 메모리 공간을 사용하기 위해서 우리는 데이터 차원을 독립적으로 다루었다. 구체적으로 설명하면 각 차원에 대해 최소값과 최대값의 범위를 구해 일정한 간격으로 나누고 가장 많은 데이터를 가진 구간에 있는 원소들을 마크한다. 모든 차원에 대해 같은 방법으로 처리한 후에 데이터 샘플들 중에 가장 많이 마크된 데이터들이 새로운 클래스의 멤버가 될 것이다. 모든 실험에서 각각의 데이터 차원을 일정하게 50등분 하고 HT-TCM-OUD에 의해 탐지된 테스트 샘플의 10분의 1만큼을 새로운 클래스 멤버로 제시하여 새로운 클래스 원소들의 탐지비율을 구해보았다. Imgchar에서 HT-TCM-OUD을 이용하여 클래스 분포변화에 대한 경고가 울린 21개의 경우, 즉 $z > z_a$= 1.64인 경우들에 대해 새로운 클래스 멤버의 탐지 비율을 측정한 결과 평균 82%의 높은 감지결과를 보여줬다. 새로운 클래스에서 1/5을 추가해서 테스트 데이터에 추가시켰더니 26개의 경우에 대해 새로운 클래스에 대해 알람을 울리고 감지비율은 95%가 되었다. 똑같은 방법으로 Isolet데이터에 대해 실험을 해본 결과, Imgchar에서 만큼 새로운 클래스 발생에 대한 경고를 울리지는 못했지만, 일단 경고가 울린 경우 새로운 클래스 멤버의 평균 감지율은 82%로 높게 나타났다. 특히 z값이 높을수록 새로운 클래스 멤버들의 감지율이 높게 나타났다. 따라서 HT-TCM에서의 가설검정은 새로운 클래스 멤버의 잘못된 탐지를 방지하는 데에서 매우 효과적이라고 할 수 있다.

[참고문헌]

[1] M. Li and I. Sethi. Confidence-based classifier design. *Pattern recognition*, 39:1230-1240, 2006.

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

[3] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition. In *Proceedings of the 13th European conference on Machine Learning*, 2002.

[4] D. Barbara, C. Domeniconi, and J. Rogers. Detecting outliers using transduction and statistical testing. *In Proceedings of the 12th ACM SIGKDD International conference on knowledge discovery and data mining*.

[5] S. Ross. *Introduction to probability and statistics for engineers and scientists*. Elsevier academic press, 2004.

[6] Uci machine learning repository.

http://www.ics.uci.edu/mlearn/mlrepository.html.