스트리밍 데이터에서의 기계학습

박정희*

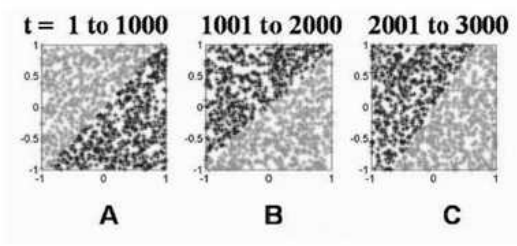
1. 소 개

대부분의 기계학습 알고리즘들은 데이터가 수집되고 저장되어진 후 이것을 한꺼번에 메모리에 올려 분석하는 배치 모드(batch mode) 방식을 기본으로 개발되어져 왔다. 그러나 점점 더 다양한 형태의 데이터 생성과 저장 방식이 가능해짐에 따라 이에 맞는 데이터 분석 방법의 개발도 꾸준히 이루어지고 있다. 시간이 흐름에 따라 끊임없이 생성되는 스트리밍 데이터(streaming data)에서는 모든 데이터를 저장하여 한꺼번에 분석하는 것이 불가능해지고 대신에 현재의 분석모델을 새롭게 생성되는 데이터의 패턴에 맞는 분석모델로 업데이트하는 게 필요하다.

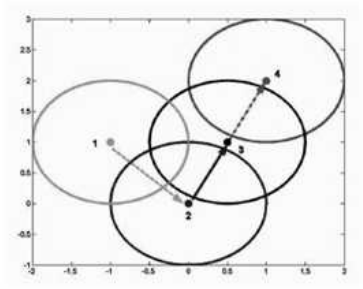
스트리밍 데이터(streaming data)는 시간에 따라 데이터 패턴이 변화되는 가능성을 내포하고 있다. 통신회사에서 통신내역을 기록하는 데이터나 전력회사에서의 전력소비와 관련된 데이터, 각종 센서 데이터 등을 예로 들 수 있다. 스트리밍 데이터에 대한 연구는 많은 연구자들의 관심을 받아 왔고 꾸준히 연구되어지고 있다. 그러나 배치모드 데이터 분석에서 높은 성능을 보이는 알고리즘들을 어떻게 스트리밍 데이터 타입에 적용할 수 있는가에 있어서는 여전히 많은 연구가 필요하다. 특히 시간이 지남에 따라 데이터 분포가 변하거나 관심개념이 달라질 수 있는 스트리밍 데이터의 특성을 고려하여야 한다[1]. 그림1(a)에서 두 클래스의 경계가 시간이 지남에 따라 변화를 보이는 예를 보이고 있다. 시간 구간 t=1~1000에서 학습된 분류 경계를 시간 t=1001 ~2000사이에서 사용한다면 경계 부근의 많은 데이터가 잘못 분류되게 될 것이다. 따라서 분류 경계가 변하고 있음을 적절한 시간 안에 탐지해내서 분류 경계 수정을 해 주는 것이 필요하다. 그림1의 (b)는 클러스터의 중심이 시간이 지남에 따라 변화되는 양상을 간단히 나타내고 있다. 타원 형태의 클러스터가 중심을 1->2->3->4의 위치로 이동하고 있다. 1의 위치에 있는 클러스터가 변하지 않고 있다고 가정하고 분석을 진행한다면 이후에 데이터 분석에서 잘못된 결과를 얻게 될 것이다. 따라서 시간이 지남에 따라 데이터 분포가 변하거나 관심개념이 달라질 수 있는 스트리밍 데이터의 특성을 고려하여 효과적으로 데이터분석을 할 수 있는 알고리즘의 개발은 매우 중요하다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 시간이 지남에 따라 데이터 분포가 변하거나 관심개념이 변경될 수 있는 스트리밍 데이터에

※ 교신저자(Corresponding Author) : 박정희, 주소 : 대전시 유성구 대학로 99 충남대학교 공대5호관 512호, 전화: 042-821-6293, FAX: 042-821-4997, E-mail : cheonghee@cnu.ac.kr



(a) 분류 경계가 변화하는 예



(b) 클러스터의 중심이 변화하는 예

그림 1. (a) 두 클래스(그레이 부분과 검은 부분)의 경계가 시간이 지남에 따라 약간씩 달라지고 있다. (b) 클러스터의 분포의 중심이 시간이 지남에 따라 1->2->3->4의 위치로 달라지고 있다.〔2〕

대한 주요 연구주제들을 살펴보고자 한다. 3절에서는 스트리밍 데이터 연구에서 이용되는 대표적인 벤치마킹 데이터 셀과 분석 툴을 알아본다. 4절에서 결론을 맺는다.

2. 스트리밍 데이터에 대한 주요 연구 분야

2.1 컨셉 변화 탐지 (concept drift detection)

컨셉 변화란 시간이 흐름에 따라 데이터 분포가 변하거나 관심있는 컨셉이 달라지는 것 등을 모두 포함한다. 데이터 발생 분포에서의 변화나 유저의 관심이 변하는 경향을 미리 탐지함으로써 분류모델이나 군집모델의 새로운 학습을 유도할 수 있다. 대표적인 변화탐지기법인 DDM(Drift Detection Method), EDDM(Early Drift Detection Method)와 EWDM(Exponentially Weighted Moving Average)는 분류기의 오류율을 이용해 변화를 탐지한다. DDM(Drift Detection Method) [3]에서는 각 데이터 샘플에 대한 예측 오류를 베르누이 시행에서의 확률 변수로 나타내고, n개의 데이터 샘플에서 발생하는 예측 오류의 수를 이항 분포로 설명한다. n개 중에 발생한 오류의 수의 비율을 나타내는 오류율의 평균에 대한 신뢰도 구간 추정을 이용하여 오류율이 급격히 증가되는 것을 탐지하고, 이를 개념 변화 발생의 시그널로 이용한다. EDDM [4]에서는 오류의 횟수 대신에 두 오류들 사이의 거리를 이용하여 점진적 변화에 대한 탐지 성능 향상을 추구하였다. EWMA [5]는 오류율에서의 급격한 변화를 탐지하기 위해 지수 가중 이동 평균을 이용하였다. 최근에는 데이터 샘플에 대한 예측이 맞을 경우에는 0으로 그렇지 않은 경우에는 1로 설정되는 이진 값으로 나타내는 오류(error) 대신에 분류 모델에 의해 얻어지는 확률 예측치를 사용하여 행동 패턴을 묘사함으로써 개념 변화 발생을 탐지할 수 있는 방법이 제안되었다[6].

오류율을 이용한 컨셉 변화 탐지 기법은 데이터 샘플들이 도착한 후에 바로 클래스 라벨 정보가 유효하게 된다는 가정 하에 수행할 수 있는 방법이다. 그러나 이러한 가정이 성립하지 않는 경우가 실제로 많이 존재한다. 클래스 라벨 정보가 제한적으로 주어지거나 시간적 딜레이를 가지고 주어질 때, 인접한 두 개의 윈도우 내의 데이터 분포가 다른지를 탐지하기 위한 통계적 테스트로

서 Wald-Wolfowitz test, Wilcoxon rank sum test, Two-sample Kolmogorov-Smirnov test 등이 사용될 수 있다. 그러나 대부분의 통계적 테스트는 1차원 데이터에서 적용가능하다. 다변수 속성을 가지는 데이터에서 통계적 테스트를 적용하기 위해서는 컨셉 변화에 대한 정보를 줄 수 있는 1차원 변수로의 변환이 이루어져야 한다.

아주 소수의 라벨드 데이터가 있을 경우 이로부터 생성된 분류모델을 이용하여 사후 확률이나 신뢰도를 이용한 1차원 변수로의 변환을 유도할 수 있다[7]. 논문 [8]에서는 SVM에 의한 분류경계의 마진 이내 영역에 있게 되는 데이터 샘플들의 밀도 변화를 이용하여 변화를 탐지하였다. 데이터 샘플들의 클래스 라벨 정보가 전혀 주어져 있지 않을 때 적용할 수 있는 컨셉 변화 탐지 방법으로 Wald-Wolfowitz test를 다변수 데이터에 대해 적용할 수 있게 확장한 방법이 있다[9].

2.2 스트리밍 데이터에서 분류 기법의 적용

미리 결정되어진 몇 개의 클래스(class) 혹은 카테고리(category) 중의 하나로 데이터 샘플들을 대응시키는 결정규칙은 이미 클래스 라벨이 알려진 학습데이터를 이용하여 구성하게 되며 다양한 분류기법(classifier)들이 패턴인식, 기계학습, 데이터마이닝 등의 분야에서 개발되어 왔다. 분류기법들을 스트리밍데이터에 적용시키는 방법은 크게 세 가지로 구분할 수 있다.

- 새로운 데이터 샘플이 들어올 때 마다 또는 새로운 데이터 샘플들의 집합에 대해 현재의 분류기를 점증적으로(incrementally) 업데이트한다. [10.11]
- 최근의 데이터 청크(chunk)에 대해 모델링된 분류기를 과거의 데이터 청크마다 학습된 분류기들과 함께 분류기 앙상블을 구성하여 예측에 적용한다. [12,13]
- 데이터 발생 분포에서의 변화나 유저의 관심이 변하는 경향을 탐지하는 변화탐지기법과 연동하여 변화탐지가 될 때마다 새로운 분류기 모델링을 시작한다.[2,3,4,5]

개념 변화 발생 또는 미발생의 이분법적인 판단 대신에 새로운 데이터 샘플이나 데이터 청크가 들어올 때마다 최근 데이터에 대한 비중을 과거 데이터에 대한 비중보다 높게 하여 분류모델을 업데이트 함으로써 개념 변화가 발생했을 때 빠르게 분류기가 적응하도록 할 수 있다. 특히 변화 발생 정도를 예측하여 최신 데이터 샘플의 비중을 자동적으로 조절하는 적응적(adaptive)인 분류 모델 업데이트 방법들이 있다[14,15].

2.3 개념변화 스트리밍 데이터에서 능동적 학습

클래스 정보가 없는 데이터(unlabeled data, 언라벨 데이터)에 비해서 클래스 정보가 있는 데이터(labeled data, 라벨 데이터)를 얻기 위해서는 비용과 시간이 들게 된다. 능동적 학습(active learning)은 소수의 라벨 데이터로 구성된 훈련집합이 주어진 경우에 분류기 학습에 가장 도움이 될 만한 언라벨드 데이터를 선택하여 전문가에 의한 라벨링을 통해 훈련 집합에 포함시키는 과정을 반복함으로써 분류기의 성능을 향상시키는 것을 목적으로 한다. 스트리밍 데이터에서 데이터 인스턴스 발생 즉시 클래스 라벨이 알려지지 않고 일정기간의 경과가 필요한 경우나 전문가에 의한 라벨링을 통해서 클래스라벨을 얻어야 할 경우, 능동적 학습을 통해 라벨링에 드는 시간과 비용을 절약할 수 있게 된다. 특히 컨셉변화 가능성이 있는 경우 정보력이 있는 적절한 인스

턴스 선택은 성능 저하를 막을 수 있는 분류기 업데이트에 중요한 요소이다.

배치 모드에서의 능동적 학습 방법들이 현재 분류기가 예측하기 가장 어려운 데이터 샘플을 선택하는 것처럼, 데이터 스트림에서의 능동적 학습 또한 예측에서의 불확실성을 이용한다. 논문 [16,17]에서는 데이터 샘플들의 분류경계에 가장 가까이 있는 샘플을 선택하거나 분류경계에서의 거리에 반비례하는 확률로 샘플들을 선택하였다. [18]에서는 언라벨드 샘플들에 대해 의사결정트리에 의한 오류를 추정하고, 오류율에서의 급격한 변화가 발생할 때 샘플들을 선택한다. [19]에서는 오류율을 이용하는 컨셉 변화 탐지 방법을 라벨드 데이터 샘플들에 대해서 적용하고, 컨셉 변화가 탐지되었을 때 샘플들을 선택하여 새로운 분류기를 구성한다.

2.4 고차원 스트리밍 데이터에서 차원 감소 기법

스트리밍 데이터 분석에서의 시간적, 공간적 제약 조건은 배치모드에서 높은 성능을 보이는 차원감소기법들의 적용에 한계를 두게 한다. 고차원데이터에 대한 성능이 우수한 차원 감소 기법들은 많이 연구되어져 온 반면, 고차원 스트리밍 데이터에서 적용할 수 있는 차원 감소 기법은 incremental PCA[20] 또는 이를 이용한 차원 감소법[21,22] 등으로 제한적이다. CCIPCA (Candid covariance-free incremental PCA)[20]는 차원감소를 스트리밍 데이터에 적용하기 위해 많이 사용되는 점증적 주성분분석 기법으로서 분산 행렬의 계산을 피하면서 주성분(principal components)들을 업데이트하는 방법이다. 스트리밍 데이터에 적용할 수 있는 선형 판별 분석(LDA) 기반 점증적 차원감소 방법들[21,22] 중에서 Incremental partial least squares(IPLS)[22] 방법은 CCIPCA를 전처리로 이용하여 분산 행렬의 근사 행렬을 구한 후 PLS를 점증적으로 업데이트하는 방법으로 좋은 분류 성능을 나타낸다고 보고되었다.

최근에 저자는 개념 변동 고차원 스트리밍 데이터에서 분류기법을 효과적으로 적용할 수 있는 방법으로서 점증적 차원 감소와 적응적 분류기를 결합시키는 방법을 제안하고, 다양한 상황에서의 실험을 통해 효과를 분석하였다[23]. 그러나 개념 변동 가능성이 있는 고차원 스트리밍 데이터에서 차원감소 기법을 어떻게 적용시켜야 분류성능 향상을 가져올 수 있는지에 대한 연구는 여전히 미비하다. 다큐먼트, 이메일, 이미지나 동영상 등 고차원 속성을 지닌 스트리밍 데이터의 발생이 보편적임을 고려할 때, 효과적으로 차원 감소 기법을 적용하여 개념 변동 고차원 스트리밍 데이터의 분석 성능을 향상시킬 수 있는 방법에 대한 연구가 필요하다.

3. 스트리밍 데이터 분석 연구에서의 실험 환경

3.1 실제 스트림 데이터 (real stream data)

컨셉 변화가 있는 데이터 스트림 연구에서 많이 활용되는 실제 데이터 셑으로 Electricity, Forest Cover type, Usenet 등이 있다. Electricity 데이터는 호주의 New South Wales주의 전기 시세를 바탕으로 전기세의 변화를 묘사한 데이터이다[24]. 45312 개의 데이터로 구성이 되어있고, 각 데이터는 8개의 속성을 가지며 클래스 라벨은 “UP”과 “DOWN” 으로 구성되어 지난 24시간 동안의 평균에 관한 가격의 변화를 정의한다. Forest Cover type 데이터는 US Forest

Service(USFS) Region 2 Resource Information System (RIS) 데이터[25]로부터 얻어진 30 X 30 m 크기의 칸의 원소들로 이루어진 forest cover type을 포함하고 있다. 데이터 목적은 지도제작 변수로부터 forest cover type을 예측하는 것이다. 581,012 개의 데이터와 54개의 속성으로 이루어져 있는데, 속성은 10개의 수치형 속성과 44개의 명목형 속성으로 구성되어 있다. Usenet 데이터는 20 newsgroup collection을 바탕으로 하는데, 시간이 지남에 따라 사용자의 관심사가 변하는 것에 관련해서 개념 변화를 설정한다[26]. 5931 개의 데이터와 659개의 속성으로 구성되어 있다.

3.2 인공데이터

급격한 또는 점진적 개념 변화와 속성의 노이즈 유무에 따라 다양한 방법으로 구성된 8개의 인공데이터가 많이 사용된다[3]. 각 데이터 셀은 한 개의 개념이 유지되는 구간에서 두 개의 클래스를 거의 같은 사이즈가 되게 일정개의 샘플을 만들고 개념 변화는 두 구간 사이의 경계에서 발생하게 한다. 또 다른 널리 사용되는 인공데이터는 스트리밍 데이터에 대한 소프트웨어 툴인 MOA[27]를 이용하여 생성할 수 있다. 여러 가지 형태의 개념변화 스트림 데이터를 생성하기 위한 다양한 스트림 제너레이터(stream generator)가 정의되어 있다. 데이터 마이닝 벤치마킹 데이터 셀들을 이용하여 개념 변화를 인위적으로 도입하여 실험에 이용하기도 한다. 예를 들어, 속성 값들의 랜덤한 permutation을 통해 데이터 분포의 변화를 만들기도 하고, 두 클래스의 데이터 샘플들이 연쇄적으로 나오도록 스트림을 이어 붙여 한 시점을 기점으로 데이터 분포가 일어난 것처럼 조정하기도 한다.

3.3 소프트웨어 툴

스트리밍 데이터 연구에 많이 사용되는 대표적인 소프트웨어로는 MOA(Massive Online Analysis)[27]를 들 수 있다. MOA는 대량의 개념변동 데이터 스트림을 다루기 위한 오픈 소스 프레임워크이다. 배치 모드 기계 학습 방법들을 구현해 놓은 WEKA(Waikato Environment for Knowledge Analysis) [28]와 연동되는 소프트웨어이다. 스트림데이터 생성 제너레이터 외에도 개념변동 스트리밍 데이터 분석을 위한 분류기들과 개념탐지 방법들도 구현되어 있다.

4. 결 론

일상생활에서 접하게 되는 많은 데이터는 스트리밍 데이터의 형태를 취하고 있다. 출퇴근길의 음악이나 영화 감상과 같은 스트리밍 서비스부터 날씨와 같은 기상데이터, 주식데이터, 전력사용 데이터, 교통 데이터, 신문기사 검색, 웹 검색 등 다양한 분야의 스트리밍 데이터에 둘러싸여 있다. 모니터링 시스템은 센서로부터 스트리밍 데이터를 생성하고 대용량의 데이터 생성과 실시간 결정 요구의 특성을 지닌다. 트래픽 매니지먼트 시스템, 보안 시스템, 서비스 모니터링 시스템 등이 그 예이다. 스트리밍 데이터에 대한 고급 분석 방법의 개발은 모든 사물이 인터넷으로 연결되어 스스로 정보를 수집하게 하고 다른 기기와 공유하며 적절한 결정까지 내리게 하는 IoT(Internet of Things) 환경에서도 반드시 필요하다. 따라서 컨셉 변화에 능동적으로 대처할 수 있는 알고리즘과 시스템의 개발은 응용의 범위가 매우 넓다.

참 고 문 헌

[1] J. Gama and I. Zliobaite, A. Bifet, M. Pechennizkiy and A. Bouchachia, "A Survey on Concept Drift Adaptation", ACM Computing Surveys, Vol. 46(4), pp. 1-37, 2014

[2] S. Ho and H. Wechsler, "A martingale framework for detecting changes in data streams by testing exchange ability," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 12, pp. 2113-2127, 2010.

[3] J. Gama. P. Medas, G. Castillo and P. Rpdrigues, Learning with drift detection, In Proceedings of SBIA Brazilian Symposium on Artificial Intelligence,2004.

[4] M. Baena-Garcia, J. Campo-Avilla, R. Fidalgo, A. Bifet, R. Gavalda, and R. Moales-Bueno. Early drift detection method, In proceedings of ECML PKDD 2006 Workshop on Knowledge Discovery from Data Streams, 2006.

[5] G. Ross, N. Adams, D. Tasoulis, and D. Hand, Exponentially weighted moving average charts for detecting concept drift, Pattern recognition letters 33(2012), pp. 191-198, 2012.

[6] 김영인, 박정희, "스트리밍 데이터에서 확률 예측치를 이용한 효과적인 개념 변화 탐지 방법", 정보과학회 논문지, vol. 43, no. 6, pp. 718-723, 2016.

[7] P. Lindstrom, B. M. Namee and S. J. Delany, "Drift detection using uncertainty distribution divergence", IEEE11th Int. Conf. Data Mining Workshops, pp. 604-608, 2011.

[8] T. S. Sethi, M. Kantardzic, "Don't pay for validation: Detecting drifts from unlabeled data using margin density", INNS Conference on Big Data, Volume 53, Pages 103 - 112, 2015.

[9] J. Friedman and L. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, Annals of Statistics, vol. 7(4), pp. 697-717, 1979.

[10] P. Domingos and G. Hulten, "Mining high-speed data streams," In Proceedings of International Conference on Knowledge Discovery and Data Mining, 2000.

[11] A. bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing", In Proceedings of SIAM International Conference on Data Mining, 2007.

[12] H. Wang, W. Fan, P. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," In Proceedings of International Conference on Knowledge Discovery and Data Mining, 2003.

[13] J. Z. Kolter and M. A. Malloof, "Dynamic weighted majority: an ensemble method for drifting concepts," Journal of Machine Learning Research, vol. 8, pp. 2755-2790, 2007.

[14] L. I. Kuncheva and C. O. Plumpton, "Adaptive learning rate for online linear discriminant classifiers," LNCS 5342, pp. 510-519, 2008.

[15] C. Anagnostopoulos, D. Tasoulis, N. Adams, N. Pavlidis and D. Hand, Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification, Statistical analysis and data mining, vol.5, pp. 139-166, 2012.

[16] P. Lindstorm, S. Delany, and B. Namee, "Handling concept drift in text data stream constrained by high labelling cost," in Proceedings of Florida artificial intelligence research society conference, 2010.

[17] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Worst-case analysis of selecive sampling for linear classification," Journal of machine learning research, vol. 7, pp. 1205 - 1230, 2006.

[18] S. Huang and Y. Dong, "An active learning system for mining timechanging data streams," Intelligent data analysis, vol. 11, pp. 401 - 419, 2007.

[19] I. Zliobatie, A. B. abd B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," IEEE transactions on neural networks and learning systems, vol. 25(1), pp. 27 - 39, 2014.

[20] Juyang Weng, Yilu Zhang, and Wey-Shiuan Hwang, "Candid covariance-free incremental principal component analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 8, pp. 1034-1040, 2003.

[21] J. Yan, B. Zhang, S. Yan, N. Liu, Q. Yang, Q. Cheng, H. Li, Z. Chen, and W. Ma, "A scalable supervised algorithm for dimensionality reduction on streaming data," Information Sciences, vol. 17, no. 6, pp. 2042-2065, 2006.

[22] X. Zeng and G. Li, "Incremental partial least squares analysis of big streaming data," Pattern Recognition, vol. 47, pp. 3726-3735, 2014.

[23] 박정희, "개념 변동 고차원 스트리밍 데이터에 대한 차원 감소 방법", 정보처리학회 논문지: 소프트웨어 및 데이터공학, 5권 8호, 2016, 게재예정.

[24] SPLICE-2 Comparative Evaluation: Electricity Pricing, Technical report UNSW-CSE-TR-9905 of The University of New South Wales, 1999.

[25] J. A. Blackard and D. J. Dean, Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables, Computers and Electronics in Agriculture 24(3) (1999), pp. 131-151, 1999.

[26]http://www.liaad.up.pt/kdus/products/datasets-for-concept-drift

[27] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis", Journal of machine learning research, vol. 11, pp. 1601-1604, 2010.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA data mining software: un update", SIGKDD Explorations Newsletter, vol. 11(1), pp. 10-18, 2009.

박 정 희

• 1998년, 연세대학교, 수학과, 이학박사
• 2004년, University of Minnesota, 컴퓨터공학과, 공학박사
• 현 재 충남대학교, 컴퓨터공학과, 교수
• 관심분야: 데이터마이닝, 패턴인식