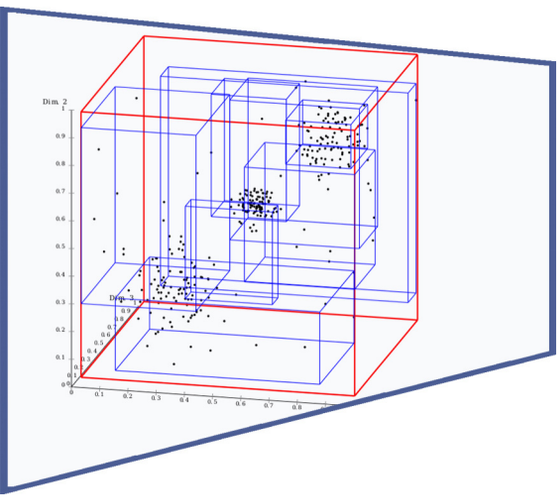

시계열 데이터 마이닝
Time Series Data Mining

1. 서론

시간영역의 측정 데이터 즉 시계열(time series data)은 현상의 분석, 경향 변화, 건전성 검토 등에 사용될 수 있는 중요한 데이터의 형태로서 다양한 분야의 과학, 공학, 경제 응용 분야에서 어렵지 않게 취득 가능하다. 이러한 시계열은 데이터의 크기, 고차원, 연속적 업데이트의 필요라는 특성으로 기술될 수 있으며 개별 값 보다는 전체적인 경향성을 분석하는 것이 필요하다. 최근 빅데이터, 데이터 마이닝 등의 대두로 인해 인공지능, 데이터 과학 분야에서 많은 연구가 수행되어 왔으며 이러한 빅데이터 기반의 분석 결과를 구조해석의 입력 데이터, 해석 방법론 개발, 최적 설계 등에 적용하기 위해 다양한 분야에서 수행된 연구를 검토하는 것이 필요하다.

본 기사에서는 특히 컴퓨터 공학, 데이터 마이닝 분야에서 수행된 시계열 분석 연구에 대해 참고문헌(Fu, 2012)의 내용을 기반으로 정리하고자 한다. 시계열 데이터 마이닝과 분석에서 중요한 요소는 시계열 데이터를 어떻게 표현하는가 하는 것으로, 주로 사용되는 방법은 시계열을 인덱싱을 통해 차원을 감소시켜 다른 영역에서 표현하는 것이다. 또한 시계열 하위 조각들에 대한 상호 유사도를 검토, 분석하는 것이 데이터 분석의 중요한 업무가 된다. 데이터 마이닝, 분석과 관련한 기존의 연구 결과들을 정리하면 데이터의 1) 표현(Representation), 2) 인덱싱(Indexing), 3) 유사도 평가 도구(Similarity measure), 4) 세분화(Segmentation), 5) 시각화(Visualization) 및 6) 시계열 마이닝(Mining in time series)로 분류할 수 있는데, 본 기사에서 이러한 연구 내용을 소개함으로서 시계열 분석과 데이터 마이닝에 대한 전반적인 이해를 높이고, 전산역학 분야의 응용에 대해 고찰하고자 한다.

2. 시계열의 표현과 인덱싱

시계열 표현의 주요한 목적은 차원 즉 표현된 데이터의 수를 감소시키는 것이다. 가장 단순한 방법은 일정한 주파수와 시간 길이를 정한 다음 데이터에서 일정 간격의 데이터를 취득하는 샘플링(Sampling)이다.

구 본 용
군산대학교 기계융합시스템공학부 조교수

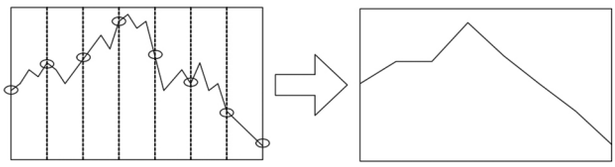

그림 1 데이터 샘플링(Sampling)

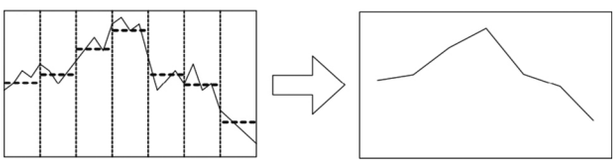

그림 2 PAA를 이용한 시계열 차원 감소

이러한 데이터 샘플링은 샘플링 주기가 너무 클 경우 데이터를 왜곡시킬 수 있으므로 일정시간 동안의 데이터 세그먼트(Segment)에 대한 평균을 사용하는 방법으로서 그림 2와 같은 Piecewise Aggregate Approximation(PAA)가 개발되었고 또한 데이터 세그먼트에 대한 시간 간격을 가변하는 Adaptive Piecewise Aggregate Approximation(APCA) 방법이 사용되고 있다.

시계열의 차원을 감소시키는 다른 방법으로서 두 가지 방법 즉 선형 보간(Interpolation)에 의한 방법과 선형 회귀(Regression)에 의한 방법을 고려할 수 있다. 선형 보간을 사용하는 일반적인 방법 중 하나는 Piecewise linear representation(PLR 또는 PLA)으로서 일정시간 동안의 시계열 $P(P_i,....,P_j)$에 대해서 처음 데이터와 마지막 데이터를 연결하는 보간 직선을 만들 수 있다. 또한 변화가 발생하는 점을 연결하는 형태로 시계열의 차원을 감소시킬 수도 있는데 Perceptually important points(PIP) 방법이라고 불린다. 패턴 매칭 및 시계열과의 최대 거리 탐색과 같은 방법을 사용하여 변화가 발생하는 변곡점을 탐지하고 이 점들을 연결하는 형태로 데이터의 차원을 줄이게 된다.

PIP 방법에서 중요한 요소는 변곡점을 어떻게 탐지할 수 있는가 하는 알고리즘으로 패턴 분석, 유사도 분석, 중요하지 않은 변동의 무시 등을 도입할 수 있다. 예를 들어 어떤 한계점 R을 정하고 일정 구간내의 하나의 값이

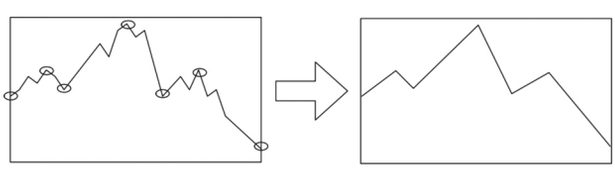

그림 3 PIP를 이용한 시계열 차원 감소

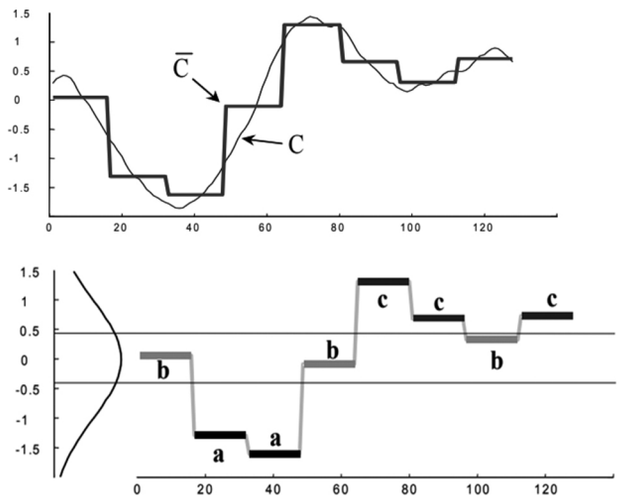

그림 4 SAX를 이용한 상징적 표현(Lin et al. 2003)

다른 값들과 비교하여 그 비율이 R 보다 크거나 작은 경우, 구간 최대값 또는 구간 최소값으로 정할 수 있다.

시계열의 차원을 감소시켜 표현하는 또 다른 방법으로 시계열을 상징적 형태(Symbol)로 바꾸어 표현하는 방법이 있다. 참고문헌(Lin et al. 2003)에서 제안된 symbolic aggregate approximation(SAX)는 PAA 방법에 의해 처리된 시계열을 상징적 표현의 문자열로 바꾸어 준다. 그림 4와 같이 PAA되 처리된 선 $\overline{C}$에 대해서 최소 분기점은 'a', 최소 분기값과 2번째 작은 분기점 사이의 분기점을 'b'에 할당하는 형태로 시계열 데이터를 'baabccbc'와 같은 상징적 형태로 바꾸어 표현하는 것이 가능하다.

지금까지는 시간 영역에서 시계열을 직접 분석하는 방법에 대해 소개하였는데 시계열을 주파수와 같은 다른 영역으로 바꾸어 분석하는 방법이 널리 사용되고 있다. 기계공학자들에게 친숙한 discrete Fourier transform(DFT)에 유사도 분석 등을 추가하여 시계열 데이터 분석에 도입되어 왔으며 유사하게 discrete wavelet transform(DWT) 방법과 시계열의 평균과 차이를 비교하는 Haar transform 또한 시계열 데이터 표현에 사용되고 있다. 또한 다중 통계분석에 주로 사용되는 principal component analysis (PCA) 방법이 금융 분야의 시계열 분석과 센서 측정 데이터 분석에 사용되고 있다. PCA는 다중 차원의 데이터를 저차원 공간으로 낮추어 주는데 서로 연관이 있는 다중 차원 데이터를 분석하고 시각화하는데 편리하다. 시스템 의특성값을 찾는 singular value decomposition(SVD) 또한 다른 형태의 transformation 기반의 방법론이 된다.

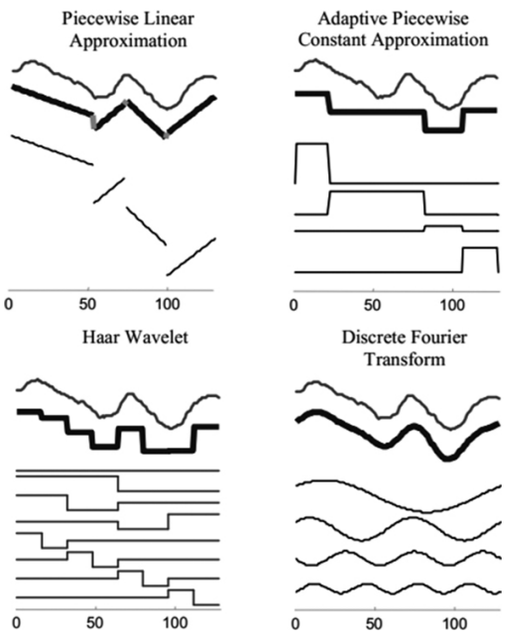

그림 5 시계열 표현의 비교(Lin et al. 2003)

앞서 설명한 APCA 방법, PLA 방법과 transformation에 기반한 두 가지 방법인 DFT과 Haar transform에 의한 표현을 그림 5에 나타내었다.

이렇게 얻어진 시계열 데이터를 표현하고 단순화된 데이터를 저장하기 위한 구조를 만들기 위해 인덱싱(Indexing)을 적용하게 되는데 대표적인 것이 그림 6의 R-tree이다. R-tree는 인접한 데이터를 그룹화 하고 그들 간의 최소거리에 대한 사각형 또는 육면체를 만드는 형태로 인덱싱을 수행한다.

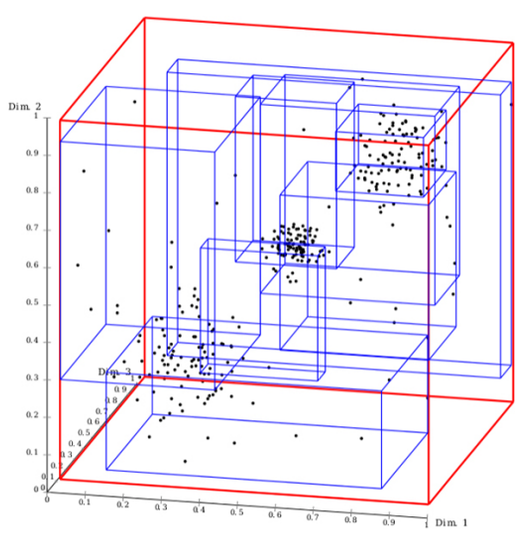

그림 6 R-tree를 이용한 인덱싱

R-tree에서 확장된 형태로 R*-tress, R+-tree, Multi-Metric tree, iSAX 등 다양한 데이터 구조가 개발되었고 앞서 설명한 데이터 표현 즉 PCA, DFT, SAX 등의 표현을 통해 변환된 데이터 구조를 정리하고 저장하는데 사용되고 있다.

3. 유사도 평가 도구

유사도 평가도구는 시계열 분석과 데이터 마이닝에서 중요한 분야이며 2장에서 소개한 많은 표현 방법 또한 유사도 평가 척도 및 도구에 대해 제안하고 있다. 전통적 데이터베이스에서 유사성 평가는 완전하게 일치하느냐 못하느냐의 문제였으나 시계열 데이터에서는 근사적인 평가를 수행하게 된다. 예를 들어 주식시장에서의 가격 변화에서 다음과 같은 질문이 가능하다.

1) 주식 1과 유사한 가격 동향 시계열을 가지는 주식을 찾으시오.
2) 지난 1달간 종가 기준 머리-어깨 패턴을 가지는 주식을 찾으시오.

이러한 질문에 답하기 위해서는 두 시계열 간의 유사도를 평가하는 도구가 필요하며 질문 1)에 대응하는 시계열 전체의 자료를 비교하는 방법(전체 유사성 확인)과 질문 2)에 대응하는 일정기간 동안의 시계열의 유사성을 찾는 도구(부분 유사성 확인)가 가능하다.

3.1 전체 유사성(whole sequence matching)

두 개의 시계열에서 유사성을 측정하기 위한 대중적인 방법은 DFT 또는 DWT에서 얻어진 성분들에 대한 차이 즉 유클리드 norm을 평가하는 것이다. 유클리드 norm을 평가 도구로 사용하는 것 외에도 변곡점, 기하학적 방법, 제한조건, 경사도 등을 평가 도구로 사용하는 것이 가능하다. 또한 시계열 블록에서 패턴 인식을 통한 유사도 검사 방법도 제안되었다.

실제로 가장 널리 사용되는 유사도 검사 도구는 “time warping”이라고 하는 거리 측정 도구이다. Dynamic time warping(DTW)에 기반한 이 방법은 속도가 다를 수 있는 두 개의 시계열 데이터의 유사도를 측장하는 검사도구로 사용될 수 있는데, 예를 들어 음성 인식, 서명 인식과 같은

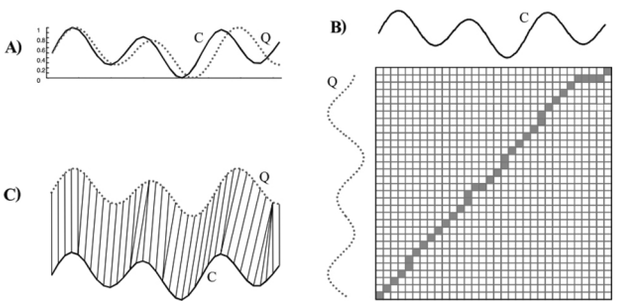

그림 7 DWT를 이용한 시계열 유사성 검토(Keogh 2002)

영상 신호 처리 사용될 수 있으며, 충격에 의한 진동 신호의 분석에도 적용된 연구가 있다.

DWT를 이용한 시계열 유사성 평가 방법을 그림 7에 나타내었다. A)에서 속도가 다른 두 시계열 C, Q에 대해서 B)와 같은 두 시계열 간의 norm을 나타내는 warping matrix W를 생성한다. 시간 별로 W의 요소를 w_k라고 하면 시작 시간과 마지막 시간을 이어주는 경로의 최소는 다음과 같이 표현된다.

$$DTW(Q,C) = \min\left\{\sqrt{\sum_{k=1}^{K} w_k}\right\}$$

위와 같이 W 행렬에서 최소 경로로 표현된 두 시계열의 차이가 속도와 무관한 두 시계열의 유사도 수치가 되게 된다. 이러한 DTW에 의한 유사도 검사 방법은 시계열의 데이터가 커질 경우 행렬 W의 크기가 매우 커지게 되어 많은 계산시간이 요구된다. 따라서 DTW의 속도를 개선하는 알고리즘이 다양하게 개발되어 왔으며, 특히 2장에서 설명한 시계열의 차원을 감소시켜 데이터를 단순화한 다음 DTW를 적용하는 연구들이 수행되어 왔다. 예를 들어 PAA, APCA와 DTW를 결합한 유사도 검사가 가능하다.

DTW 유사도 검사 도구와 유사한 주제에 대해서 Longest common subsequence(LCSS) 방법과 Threshold-based distance function 방법, parameter-light distance measure 방법 등이 제안되어 왔다.

3.2 부분 유사성(sequence matching)

부분 유사성 검사는 일정 시간의 시계열이 주어졌을 경우 이 시계열과 기존에 존재하는 긴 시간 동안의 시계열을 비교하여 유사성을 찾아내는 검사 도구이다. 이러한 부분 유사성 검토는 어떠한 검사도구 즉 DFT 또는 DWT를 일정 시간 시계열에 적용하고 이 검사도구의 효율적인 수행하는 알고리즘 개발에 집중되어 왔다. 이러한 방법으로는 시계열 윈도우 크기의 영향을 제거하는 GeneralMatch, DualMatch 등이 있으며 SAX, PLR과 같은 자료 구조의 인덱싱을 도입하여 계산 및 유사성 검토 속도를 높이는 시도가 있어 왔다. 또한 일정 시간 시계열에 DWT를 도입한 알고리즘이 제안되었다.

4. 시계열 마이닝

시계열 데이터 마이닝의 최종적인 목적은 시계열 분석을 통해 숨겨진 정보 또는 지식을 발견하는 것으로 시계열에 숨겨진 패턴 분석과 패턴의 분류에 적용될 수 있다.

4.1 패턴 분석과 클러스터링

시계열 데이터에서 특정한 신호와 이상 징후 등을 확인하는 것이 패턴 분석의 목적이며 이러한 패턴 분석을 위해 비슷한 시계열 데이터를 묶어주는 클러스터링을 적용할 수 있다. 클러스터 분석 자체가 특정한 알고리즘이라기보다는 해결하여야 하는 문제로 생각할 수 있으며, 문제의 이해에 따라 다양한 알고리즘이 가능하다. 예를 들어 각 멤버간의 거리, 간격, 통계 분포에 따라 그룹을 정할 수 있고, 일종의 다목적 최적화 문제로 공식화할 수 있다.

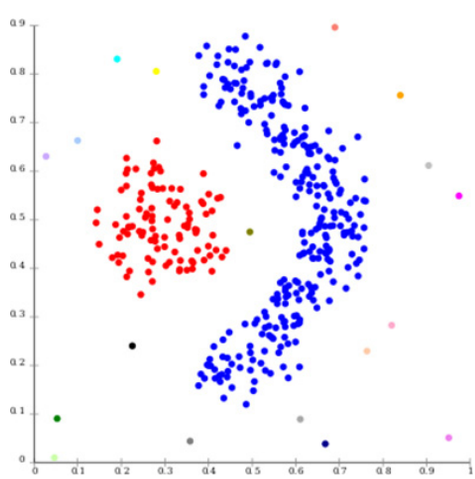

그림 8 밀도 기반 클러스터링

대표적 시계열 분석 모델인 ARMA와 ARIMA를 도입하여 데이터를 클러스터링할 수 있는데 ARMA는 직전의 데이터가 현재의 데이터에 영향을 주는 Autocorrelation 모델과 이동평균 모델인 MA를 결합한 것이며, ARIMA의 경우 ARMA 모델에 과거 데이터의 추세까지 반영한 모델을 사용한 것이다. 또한 은닉 마르코프 모델을 시계열에 적용하여 클러스터링을 수행한 연구들이 있다.

데이터를 하나의 다수의 클러스터 그룹에 할당하는 Fuzzy 클러스터링 알고리즘(FCM)이 짧은 시간과 고르지 않게 분포된 시계열의 클러스터링에 사용되었다.

4.2 분류

데이터의 분류는 전통적인 데이터 마이닝의 업무로서 이러한 데이터의 분류를 위해 시계열의 특성에 따른 패턴을 분석, wavelet 변환을 통한 요소 분류, metafeature 방법 등이 적용되어 왔다. 시계열에 대해 분류하는 인자를 정하는 연구 또한 수행되었는데 동역학 모델링이나 DTW 결정 트리에 기반한 결정 방법이 제안되었다.

5. 맺는 말

시계열을 분석하고 해석하는 것은 시계열로부터 어떤 법칙을 만들고 이를 통해 미래를 예측하기 위한 것이다. 이러한 시계열은 생활 속의 다양한 통계량, 주가의 변화, 계측기의 측정값 등으로부터 획득 가능한데 데이터 과학의 발전, 빅데이터의 대두와 함께 다양한 분석 도구 들이 개발되어 온 것을 알 수 있다.

전산 구조역학 또한 주어진 자료와 계측기의 측정치 등의 입력으로부터 요구 사항을 만족하는 최적의 구조 형태를 찾는 것이 목적이며, 또한 유한 요소 해석의 기본 방법론이 가정된 근사 변위에 대한 변형률 에너지를 최소화하는 형태로 근사해를 결정하는 것이므로 일정 부분 학문 간의 유사성을 생각해 볼 수도 있다.

구조물에 부착된 센서로부터 계측된 시계열 데이터의 분석과 이를 통한 구조 건전성 확인과 해석 모델의 수정 등은 지금도 활발하게 수행되고 있는 연구 분야이며, 향후 데이터 마이닝과 결합한 전산 구조공학 연구 등을 통해 학제 간 융합적인 연구가 늘어나는 것을 기대해 본다.

참고문헌

Fu. T.-C. (2011) A review on time series data mining, Eng. Appl. Appl. Artif. Intell. 24, pp. 164-181.

Lin. J., Keogh, E. Lonardi, S., Chiu, B. (2003) A Symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the Eighth ACM SIGMOD International Conference on Management of Data Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2-11.

Keogh, E. (2002) Exact indexing of dynamic time warping. In: Proceedings of the 28th International Conference on Very Large Databases, pp. 406-417.