의사결정나무와 유전알고리즘으로 잔존수명을 예측하기 위한 시계열 shapelet 선택방법

황석범*, 진민기*, 안길승**,허 선*
*한양대학교 산업경영공학과
**(주)현대자동차
e-mail:hursun@hanyang.ac.kr

Shapelet Selection based on Genetic Algorithm for Remaining Useful Life Prediction with Decision Tree

Seok-Beom Hwang*, Min-ki Jin*, Gil-Seung Ahn**, Sun Hur*
*Dept. of Industrial & Management Engineering, Hanyang University
**Hyundai Motors Inc.

요 약

반도체 장비나 자동차, 비행기 등 유지보수에 많은 비용이 들고 내부 시스템이 복잡해짐에 따라 잔존 수명 예측이 더욱 중요해지고 있다. 본 논문은 잔존수명 예측을 위한 시계열 데이터의 일부분인 Shapelet을 잔존수명 예측에 맞게 재정의하고 의사결정나무와 유전알고리즘에 의해 잔존수명 예측에 적용하는 방법을 제안한다.

1. 서론

기계 시스템의 잔존 수명(remaining useful life, RUL)은 현재 시점부터 고장 시점까지의 길이 또는 비율이다. RUL 예측은 고장 진단과 건전성 관리 분야에서 매우 중요한 과제이고, 예측된 RUL을 사용해 유지보수 스케쥴링을 한다. RUL을 예측하는 접근은 물리기반 모형과 데이터기반 모형으로 나눈다. 물리기반 모형은 시스템 고장 메커니즘과 같은 영역 지식을 사용하여 RUL을 예측한다. 데이터기반 모형은 이전에 관측된 데이터로부터 퇴보 패턴을 발견하고 통계 모형, 기계학습 모형, 심층학습 모형 등을 사용해 RUL을 예측한다. 본 논문에서는 기계 학습을 사용하는 데이터 기반 모형에 초점을 맞춘다.

2. 연구배경

2.1 유전 알고리즘

유전 알고리즘은 특징 선택, 사용자 매개변수 튜닝, Shapelet 선택과 같은 시계열 데이터 분석 문제에서 흔히 쓰이는 메타 휴리스틱 기법 중 하나이다. 첫 번째, 부모 집단이라고 불리우는 해집합은 무작위로 초기화된다. 두 번째, 현재의 부모 집단은 적합도 함수를 사용하여 평가되고 그 중 높은 적합도를 지닌 해들이 선택된다. 세 번째, 선택된 해들에 교차 연산과 돌연변이 연산을 진행하여 자식 해를 만든다. 여기서 교차 연산을 통해 두 개의 부모 해를 무작위로 선택하여 섞어서 자식 해를 만든다. 돌연변이 연산은 자식 해에 변동을 주어 부모 집단에 있는 해와 겹치지 않는 자식 해를 만드는 것이 목적이다. 네 번, 자식 해와 선택된 해들로 새로운 해 집단을 구성한다. 종료 조건이 만족된다면 가장 높은 적합도를 지닌 해가 최종 선택되고, 종료 조건을 만족시키지 않는다면 만족할 때까지 (2)단계부터 (4)단계까지 반복한다.

2.2 Shapelet

Shapelet은 시계열 분류 문제에서 개별 부류까지의 거리가 부류 관련성을 최대화하는 부분 시계열이다. 즉, Shapelet은 각 부류와의 거리를 다음과 같이 정의할 때 분류기의 손실 함수를 최소화하는 부분 시계열이다.

$$s = \arg\ L(f(d(s', X)), Y) \qquad (1)$$

$d(\cdot)$는 유클리디안(Euclidean)거리이고 S는 모든 가능한 부분 시계열, X와 Y는 시계열의 개체(instance)와 정답(label)을 의미한다. 마지막으로 $L(\cdot)$은 분류기의 손실함수를 의미한다. 탐색 공간 S가 너무 커서 (1)에서의 S를 쉽게 찾을 수 없기 때문에 여러가지 가정을 세워 Shapelet을 추정하기 위해 많은 휴리스틱 접근방법들이 제안되었다. 여기서 추정은 손

실 함수를 최소화하는 것이 아니며 또한 데이터셋의 부분 시계열을 찾는 것이 아닌 짧은 시계열을 찾음으로써 해결할 수 있다.

2.3 결정 나무

결정 나무는 지도학습으로 분류규칙을 정해 데이터를 분류하는 대표적인 분류기이다. 분기가 되는 기준은 불순도를 기준으로 하는데 분류가 이루어졌을 때 서로 다른 부류를 가진 데이터들이 많으면 불순도가 올라가는 원리를 사용한다. 불순도는 보통 지니(Gini) 계수와 엔트로피(entropy) 계수로 나뉜다. 엔트로피를 이용해 분기 이전의 불순도와 이후의 불순도 차이를 정보획득(Information Gain)이라고 하는데 이 정보획득을 최대로 하는 방향으로 분기한다.

3. 기존 연구

Malinowski *et al*(2015)은 처음으로 RUL shapelet을 제시하였다. 이는 잔존 수명을 예측하기 위한 특징 벡터로 사용되는 부분 시계열의 거리이다. Malinowski *et al*(2015)에서 잔존 수명을 예측하는 방법은 다음과 같다. ω는 데이터의 전체 시계열, T_i는 ω에 포함된 표본 시계열이고 표본 시계열들의 길이는 모두 다르다. $t_1, \cdots, t_{|T_i|}$는 각 시계열 를 구성하는 부분 시계열이다. S는 shapelet이며 $s_1, \cdots, s_{|S|}$는 각 shapelet을 구성하는 하나의 특징이다.

우선 RUL shapelet을 추출하기 위해 *k-means* 군집화를 사용해 추출하고 각 군집의 중심점을 RUL shapelet으로 설정한다. RUL shapelet의 수는 사용자 파라미터로 결정한다. 선택된 RUL shapelet을 표본별로 유클리디안 거리를 계산해 최소인 거리(d)를 찾고 이 최소거리가 나타난 지점에서의 남은 수명(ρ)을 찾는다.

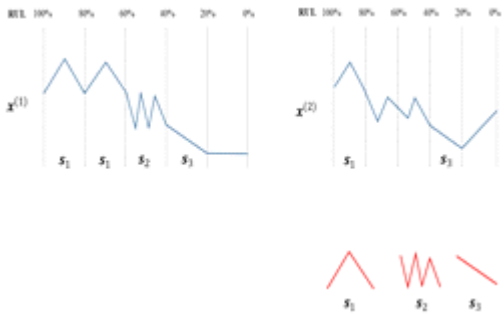
하나의 RUL shapelet에 대해 표본별로 모두 계산한 최소 거리와 남은 수명을 저장하고 최소 거리를 오름차순으로 정렬한다. 최소 거리로 정렬된 남은 수명을 정규화하여 최소 거리가 가장 작은 거리에 대응되는 남은 수명을 기준으로 순서대로 남은 수명을 하나씩 추가해 분산이 최소가 되는 색인(index)을 찾는다. 해당 색인에 대응되는 거리를 δ, 해당 색인까지 남은 수명의 평균 μ를 구한다. 하나의 RUL shapelet에 적용한 과정을 모든 RUL shapelet에 적용해 각각의 거리(δ)와 남은 평균 수명(μ)을 구해준다.

앞에서 구한 RUL shapelet과 δ, μ를 이용해 테스트 데이터(U)에 적용한다. 적용 방법은 테스트 데이터를 각 RUL shapelet과의 유클리디언 거리를 계산한다. 만약 계산한 거리가 RUL shapelet에 해당하는 δ보다 작다면 대응(matching)된 것이고 모든 남은 평균 수명(μ)과 이때의 테스트 데이터의 색인을 평균해 테스트 데이터의 전체 수명을 예측하는 방식이다.

4. 제안 방법

4.1 좋은 잔존 수명 쉐이플릿의 특징

여기서는 본 방법론의 근간이 되는 좋은 잔존수명 쉐이플릿(RUL shapelet)의 특징을 설명한다. 좋은 RUL shapelet은 모든 시계열 샘플의 비슷한 위치에서 나타나야 하고 다른 위치에서 나타나면 안된다. [그림 1]을 보면 시계열 $x^{(1)}$과 $x^{(2)}$의 좋고 나쁜 RUL shapelet, 그리고 RUL shapelet S_1, S_2, S_3이 나타나 있다. 이 그림에서는 S_1, S_2가 나쁜 RUL shapelet이고 S_3이 좋은 RUL shapelet이라고 할 수 있는데, 왜냐하면 S_1은 $x^{(1)}$과 $x^{(2)}$에서 나타나지만 [100%, 80%] 구간과 [80%, 60%] 두 구간에서 나타나기 때문이다. 이에 반해 S_3은 과 $x^{(1)}$과 $x^{(2)}$에서 오직 [40%, 20%]구간에서만 나타난다.



[그림 1] 좋고 나쁜 RUL shapelet 예시

본 논문에서는 좋은 RUL shapelet을 좋은 RUL shapelet 집합으로 확장한다. 동일 구간에서 좋은 RUL shapelet이 두 개 이상 나오면 중복성이 생겨 지도 학습 모델의 예측력을 저해시킬 수 있다. 이와 반대로 다른 구간에서 발생하는 좋은 RUL shapelet. 즉, 서로 중복되지 않는 RUL shapelet은 모델의 정확도를 향상시킬 수 있다.

4.2 초기화

4.1절에서 설명했듯이 좋은 RUL shapelet 집합이 되기 위한 조건은 다음과 같이 요약할 수 있다. 첫 번째, 개별 RUL shapelet은 RUL과 높은 상관관계를 지닌다. 두 번째, 모든 RUL shapelet은 대부분의 시계열의 특정 구간에서 나타난다. 세 번째, 동일 구간에서 발생하는 RUL shapelet은 한 개면 된다. 이 조건에 따라 본 연구에서는 유전 알고리즘의 초기화 방법을 개발한다.

$S=\{s_1, \cdots, s_m\}$를 RUL shapelet 집합 즉, 본 알고리즘의 해라고 정의한다. 훈련 데이터셋 $D=\{x^{(i)}, y^{(i)} | i=1,2,\cdots n\}$를 사용한 S의 진화과정은 첫 번째, RUL shapelet의 m수 은 이산 균등 분포에서 추출되며 ($DU(2,M)$ RUL shapelet의 최대 수를 나타내는 사용자 매개변수이다. 두 번째, [0, 1]은 무작위로 m개의 구간으로 나뉘며 다음과 같은 과정을 따른다. (1) 각 구간은 $1/m, 2/m, \cdots, (m-1)/m$로 동일한 길이를 갖는다. (2) 각 구간에 노이즈 e_j를 주고 e_j는 연속균등분포 $CU(-0.1, 0.1)$를 따른다. (3) 분기점을 근간으로 다음과 같이 구간 집합 을 생성한다.

$$V=\left\{\begin{array}{l}[(m-1)/m+e_{m-1}, 1], \\ [(m-2)/m+e_{m-2}, (m-1)/m+e_{m-1}), \cdots, \\ [0, 1/m+e_1)\end{array}\right\}$$

세 번째, 모든 i에 대하여 $(x^{(i)}, y^{(i)})$를 구간으로 나눈다. 네 번째, 각 구간에 대하여 $2 \leq k \leq (e_i - s_i)$길이를 갖는 부분 시계열의 중심 c_k를 다음과 같이 계산한다.

$$c_k = (1/n) \times \sum_{i=1}^{n} \frac{\sum_{t=s_i}^{e_i-k} x_{t:t+k}^{(i)}}{e_i - s_i}$$

마지막으로, 시계열과 중심점과의 거리가 가깝고 동시에 RUL과 상관관계가 높은 중심점을 선택하고 이를 식 다음 식에 나타낸 것과 같이 모든 에 대하여 해 s_j로 사용한다.

$$s_j = \rho((d(x_{1:t}^{(i)}, c_{k)})_{i,t}, (y_t^{(i)})_{i,t})$$

참고문헌

[1] Malinowski, S., Chebel-Morello, B., and Zerhouni, N. (2015). Remaining useful life estimation based on discriminating shapelet extraction. *Reliability engineering & system safety*, 142, 279-288

[2] Ye, L., and Keogh, E. (2009). Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 947-956).

[3] Vandewiele, G., Ongenae, F., and Turck, F. D. (2021). GENDIS: Genetic Discovery of Shapelets. *Sensors*, 21(4), 1059.

[4] Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4(2), 65-85.

[5] Si, X. S., Wang, W., Hu, C. H., and Zhou, D. H. (2011). Remaining useful life estimation-a review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1), 1-14.